EST-46115: Modelación Bayesiana.

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Toma de decisiones.
Objetivo: En esta sección conectaremos conceptos que hemos visto durante el curso para toma de decisiones bajo incertidumbre. Anteriormente, introdujimos el concepto de inferencia Bayesiana dentro del marco de teoría de la decisión. Retomaremos ciertos componentes de este marco en el contexto de inferencia y modelado predictivo..
Lectura recomendada: Sección 9.9 de [3]. Capítulo 9 de [1]. Algunos pasajes de [5]. el libro de Gilboa [2] es una excelente referencia para el modelado de toma de decisiones bajo incertidumbre.

1. INTRODUCCIÓN

Retomaremos la discusión sobre toma de decisiones bajo incertidumbre. Para tomar decisiones debemos de especificar la noción de `utilidad` asociada a cada opción que podemos tomar. La decisión `óptima Bayesiana` corresponde a la que maximiza la utilidad esperada. Buscaremos ejemplificar cómo podemos utilizar `Stan` para estimar la distribución de posibles respuestas bajo decisiones y calcular utilidades esperadas.

2. DEFINICIÓN DE UN PROBLEMA DE DECISIÓN

Siguiendo a [1] un problema de decisión Bayesiano necesita de los siguientes componentes:

1. Definir un conjunto de posibles resultados X.
2. Definir un conjunto de posibles decisiones D.
3. Definir una función de utilidad que contraste la decisión, d, contra el resultado x, la cual denotamos por $U(x, d)$.
4. Especificar nuestro estado de conocimiento sobre las posibles realizaciones de posibles resultados a través de $\pi(x)$.

También es usual, como veremos adelante, que consideremos una $\pi(x)$ para cada decisión que podamos tomar, lo cual lo denotamos por $\pi(x|d)$ si fuera necesario.

Con lo cual podemos escoger la decisión d^* que obtenga la mejor utilidad esperada

$$d^* = \arg\max_d \bar{U}[d]\,, \tag{1}$$

donde

$$\bar{U}[d] = \mathbb{E}[U(X, d)] = \int U(x, d)\pi(x)\mathrm{d}x\,. \tag{2}$$

Los resultados deberán de representar la mayor información posible que sea relevante para la especificación de la función de utilidad.

2.1. Contexto bayesiano

- Las decisiones pueden ser sobre los parámetros del modelo, θ, o cantidades observables, $\tilde{y}$.
- Nuestro estado de conocimiento lo definimos como la distribución posterior o predictiva posterior.
- La función de utilidad depende del contexto del problema.

2.2. Caso: enfoque predictivo

Supongamos que tenemos el siguiente problema de decisión bayesiano con un `enfoque predictivo`:

- Los **estados** posibles son `cantidades observables` $\tilde{y}$.
- Las **decisiones** que podemos tomar son sobre *todas* los posibles `funciones de probabilidad` relevantes para nuestro problema. Es decir, cualquier $d(\tilde{y})$ donde d es una distribución de probabilidad sobre cantidades observables.
- La **función de utilidad** que escogeremos será `utilidad logarítmica`, $\log(d)$.
- Nuestro **estado de conocimiento** sobre los estados inciertos lo reflejamos a través de la `distribución predictiva posterior`, $\pi(\tilde{y}|\underline{y}_n)$.

La decisión que maximiza la utilidad esperada bajo nuestro estado de conocimiento será la que `maximice`

$$\int \log d(\tilde{y})\, \pi(\tilde{y}|\underline{y}_n)\, \mathrm{d}\tilde{y}\,. \tag{3}$$

Nota que pedimos que sea la que `minimice`

$$-\int \log d(\tilde{y})\, \pi(\tilde{y}|\underline{y}_n)\, \mathrm{d}\tilde{y}\,, \tag{4}$$

que es justamente la `entropía cruzada` entre dos distribuciones y que sabemos tiene un punto óptimo siempre y cuando utilicemos la misma distribución con la que reflejamos nuestro estado de conocimiento.

Es decir, bajo un `enfoque predictivo` nuestra mejor decisión bajo utilidad logarítmica es utilizar la densidad predictiva posterior.

2.3. Caso: enfoque de inferencia

Supongamos que tenemos el siguiente problema de decisión bayesiano con un `enfoque de inferencia`:

- Los **estados** posibles son `la configuración del modelo` que especifica un modelo probabilístico, θ.
- Las **decisiones** que podemos tomar son *todas* los posibles `funciones de probabilidad` relevantes para nuestro poblema. Es decir, cualquier $d(\theta)$ donde d es una distribución sobre configuraciones de un modelo.
- La **función de utilidad** que escogeremos será `utilidad logarítmica`, $\log(d)$.
- Nuestro **estado de conocimiento** sobre los estados inciertos lo reflejamos a través de la `distribución posterior`, $\pi(\theta|\underline{y}_n)$.

2.3.1. Pregunta: ¿Cuál será la mejor decisión que podemos tomar en este escenario?

3. ANÁLISIS DE DECISIÓN

Vamos a seguir el ejemplo que está en la documentación de `Stan`. En este escenario el tomador de decisiones tiene que decidir el medio de transporte para llegar a su trabajo: caminar, bicicleta, transporte público o taxi.

A lo largo del año ha registrado 200 días de trayectos a su trabajo y ha registrado el tiempo que le toma llegar.

3.1. Definición de `decisiones` y `observaciones`

- Las **decisiones** son el medio de transporte codificadas numéricamente.
- Los **resultados** $X = \mathbb{R} \times \mathbb{R}$ que observamos son el tiempo t que toma y el costo c asociado a ese tiempo, $x = (c, t)$.

3.2. Definición de `estado de conocimiento`

Necesitamos definir $\pi(x|d)$ la distribución de resultados posibles sujeta a la decisión que se ha tomado. Bajo el enfoque Bayesiano ésta será la distribución predictiva posterior de una observación condicional en la historia que hemos visto

$$\pi(\tilde{x}|d, \underline{x}_n, \underline{d}_n) = \int \pi(\tilde{x}|d, \theta)\, \pi(\theta|\underline{x}_n, \underline{d}_n)\, \mathrm{d}\theta\,. \tag{5}$$

Por simplicidad utilizamos una distribución log-normal para los tiempos de llegada bajo cada transporte. Es decir, para una observación $x_n = (c_n, t_n)$ asociada a la decisión d_n consideramos

$$t_n \sim \mathsf{LogNormal}\left(\mu_{[d_n]}, \sigma_{[d_n]}\right)\,. \tag{6}$$

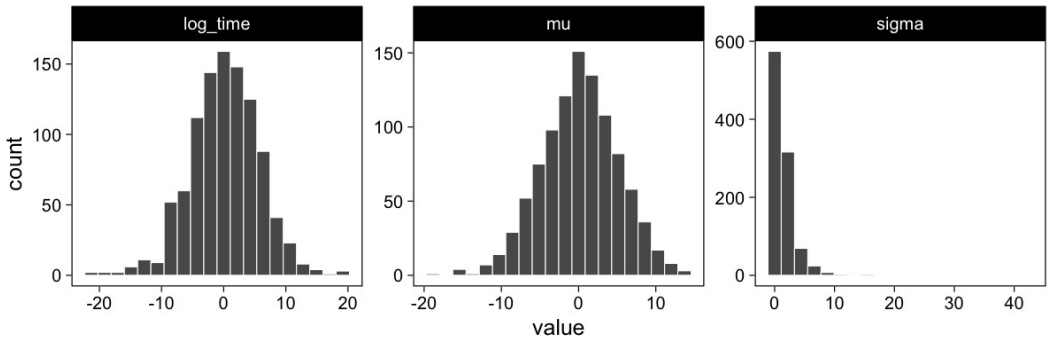
$$c_n \sim \mathsf{LogNormal}\left(\nu_{[d_n]}, \tau_{[d_n]}\right)\,. \tag{7}$$

> Decimos que una variable aleatoria se distribuye log-normal, denotado como $Y \sim \mathsf{logNormal}(\mu, \sigma)$, si $\log Y \sim \mathsf{Normal}(\mu, \sigma)$.

Las previas que utilizamos para el tiempo de llegada en cada modo de transporte, $k \in \{1, \ldots, 4\}$, son

$$\mu_k \sim \mathsf{Normal}(0, 5)\,, \tag{8}$$

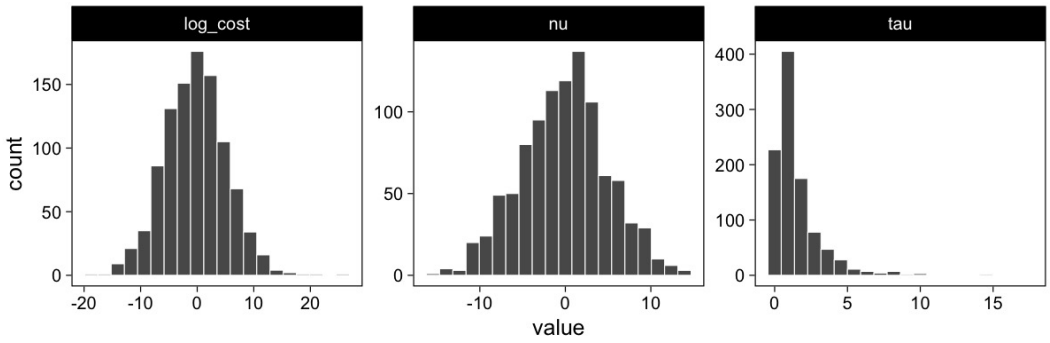
$$\sigma_k \sim \mathsf{logNormal}(0, 1)\,. \tag{9}$$



Las previas que utilizamos para los costos por cada modo de transporte, $k \in \{1, \ldots, 4\}$, son

$$\nu_k \sim \mathsf{Normal}(0, 5)\,, \tag{10}$$

$$\tau_k \sim \mathsf{logNormal}(0, 1)\,. \tag{11}$$



El conjunto de parámetros del modelo que marginalizará en la predictiva posterior es

$$\theta = (\mu_{1:4}, \sigma_{1:4}, \nu_{1:4}, \tau_{1:4})\,. \tag{12}$$

3.3. Definición `función de utilidad`

Digamos que el tomador de decisión evalúa su tiempo de traslado de manera lineal y que el tiempo invertido en transporte lo evalúa en $25 por cada momento que éste pasa en su trayecto, por lo que la función de utilidad es

$$U(c, t) = -(c + 25 \cdot t)\,. \tag{13}$$

> Nota que podríamos considerar una utilidad distinta para cada modo de transporte, $U(x, d)$, de tal manera que se reflejen costos individuales de cada medio de transporte.

3.4. Cálculo de utilidad esperada

Lo que necesitamos ahora es poder calcular la utilidad esperada de cada una de las posibles decisiones y tomar la que minimice dicha función. El siguiente código aprovecha que nuestro espacio de posibles decisiones es pequeño.

```
functions {
  real U(real c, real t) {
    return -(c + 25 * t);
  }
}
data {
  int<lower=0> N;
  array[N] int<lower=1, upper=4> d;
  array[N] real c;
  array[N] real<lower=0> t;
}
parameters {
  vector[4] mu;
  vector<lower=0>[4] sigma;
  array[4] real nu;
  array[4] real<lower=0> tau;
}
model {
  mu ~ normal(0, 1);
  sigma ~ lognormal(0, 0.25);
  nu ~ normal(0, 20);
  tau ~ lognormal(0, 0.25);
  t ~ lognormal(mu[d], sigma[d]);
```

```
  c ~ lognormal(nu[d], tau[d]);
}
generated quantities {
  array[4] real util;
  for (k in 1:4) {
    util[k] = U(lognormal_rng(mu[k], sigma[k]),
                lognormal_rng(nu[k], tau[k]));
  }
}
```

Lo que esta calculando `Stan` son los términos para estimar la utilidad esperada por medio de un `estimador Monte Carlo`. Esto lo vemos de la expresión

$$\bar{U}[d] = \mathbb{E}[U(X,d)|\underline{x}_n, \underline{d}_n] = \int U(x,d) \cdot \pi(x|d,\theta) \cdot \pi(\theta|\underline{x}_n, \underline{d}_n)\, \mathrm{d}\theta\, \mathrm{d}x\,, \tag{14}$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} U(x^{(m)})\,, \tag{15}$$

donde

$$x^{(m)} \sim \pi(x|d,\theta^{(m)})\,, \tag{16}$$

$$\theta^{(m)} \sim \pi(\theta|\underline{x}_n, \underline{d}_n)\,. \tag{17}$$

4. FUNCIONES DE UTILIDAD

La función de utilidad depende de la aplicación. En particular, de las características del problema y de la pregunta que se requiere responder con el análisis.

Bajo el contexto de `modelado predictivo` vimos que hace sentido utilizar la log-densidad predictiva posterior para tomar decisiones. Esto es cuando queremos escoger un modelo probabilístico con buenas capacidades predictivas.

Sin embargo, en algunas aplicaciones nos puede interesar hacer predicciones puntuales. Por ejemplo, en una aplicación nos puede interesar utilizar el concepto de `pérdida cuadrática` para tomar decisiones. La función de utilidad la podemos definir como

$$U_{\mathsf{Q}}(\tilde{y}, d) = -(\tilde{y} - d)^2\,. \tag{18}$$

De manera análoga, podemos utilizar nuestras nociones de `pérdida en valor absoluto` o `pérdida en valor absoluto relativo` para definir utilidades. Incluso podemos utilizar `pérdidas por intervalos` o `pérdidas por predicción correcta/incorrecta` (pérdida 1-0).

En la literatura existen algunos ejemplos de funciones de utilidad como en la Sección 9.9.1 de [3] o la Sección 9.4 de [1]. Los libros [2] o [4] proveen de un muy buen marco teórico para estos conceptos.

4.1. Ejemplo: Prueba Bechdel

Consideremos nuestro ejemplo del curso sobre las películas que pasan la prueba de Bechdel. Consideremos nuestra previa como una $\mathsf{Beta}(5, 11)$. También consideremos análisis predictivo bajo el enfoque de funciones de pérdida para las predicciones:

1. Utilidad cuadrática;
2. Utilidad 1-0.
3. Utilidad por intervalos.

El código que utilizaremos en `Stan` es el siguiente:

```
functions {
  real quadraticUtility(int y_tilde, int d) {
    return -(y_tilde - d)^2;
  }
  real zeroOneUtility(int y_tilde, int d){
    if (y_tilde == d) {
      return 1;
    } else {
      return 0;
    }
  }
  real intervalUtility(int y_tilde, int d){
    if (fabs(y_tilde - d) < 10) {
      return 0;
    } else {
      return -fabs(y_tilde - d);
    }
  }
}

data {
  int<lower=0> N;
  int<lower=0> K;
  array[N] int<lower=0, upper=1> test;
}

parameters {
  real<lower=0, upper=1> theta;
}

model {
  theta ~ beta(5, 11);
  test ~ bernoulli(theta);
}

generated quantities {
  array[K] real utilQuad;
  array[K] real utilZeroOne;
  array[K] real utilInterval;
  for (kk in 1:K) {
    utilQuad[kk] = quadraticUtility(binomial_rng(K, theta), kk);
    utilZeroOne[kk] = zeroOneUtility(binomial_rng(K, theta), kk);
    utilInterval[kk] = intervalUtility(binomial_rng(K, theta), kk);
  }
}
```

```
posterior ← modelo$sample(data = c(N = nrow(data),
                                   K = 20,
                                   data),
                          refresh = 10000,
                          iter_sampling = 4000,
                          seed = 108727)
```

Los resultados se muestran a continuación de manera gráfica (Fig. 1).

Notas que bajo pérdida cuadrática y pérdida 1-0 podemos identificar un único punto máximo de utilidad esperada. Bajo pérdida por intervalos, no.

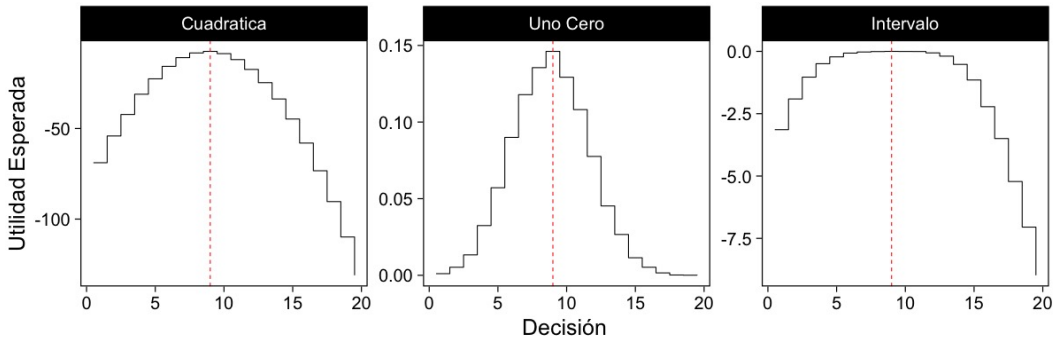


FIGURA 1. *Utilidad esperada bajo distintas funciones de utilidad.*

Bajo ciertas funciones de utilidad podemos identificar los resúmenes adecuados que maximicen la función de utilidad. Por ejemplo, para pérdida cuadrática corresponde el valor esperado posterior (en el ejemplo de las películas el de la distribución predictiva posterior). Para pérdida 1-0, la moda.

5. DECISIONES CONTINUAS

El ejemplo anterior utilizaba decisiones discretas (o un espacio de decisiones discretas). Si las decisiones fueran sobre un continuo, el problema se vuelve mas complicado, para lo cual las capacidades actuales de `Stan` son limitadas.

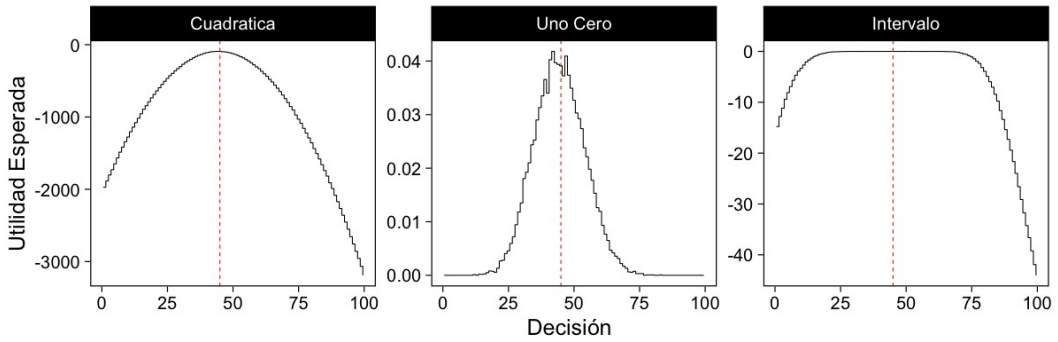


FIGURA 2. *Utilidad esperada bajo distintas funciones de utilidad.*

REFERENCIAS

[1] 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, 5

[2] I. Gilboa. Theory of Decision under Uncertainty. 1, 5

[3] O. A. Martin, R. Kumar, and J. Lao. *Bayesian Modeling and Computation in Python*. Chapman and Hall/CRC, Boca Raton, First edition, 2021. 1, 5

[4] J. Q. Smith. *Bayesian Decision Analysis: Principles and Practice*. Cambridge University Press, Cambridge, 2010. ISBN 978-0-521-76454-4. . 5

[5] A. Vehtari and J. Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6(none):142–228, jan 2012. ISSN 1935-7516. . 1