Modelos Bayesianos con R y STAN

Andrés Gutiérrez - Hanwen Zhang

2021-05-29

Índice general

Capítulo 1

Prefacio

θ

Π

Capítulo 2

Tópicos básicos

2.1 Teoría de la decisión

El problema estadístico de estimar un parámetro se puede ver dentro del contexto de la teoría de decisión: la estimación que proveemos, sea en el ámbito de la estadística clásica o la estadística bayesiana, depende de los datos muestrales, $\mathbf{X}$, de tal forma que si éstos cambian, nuestra estimación también cambia. De esta manera, el proceso de estimación puede ser representado como una función que toma un conjunto de datos muestrales y los convierte en una estimación de nuestro parámetro de interés, $A(\mathbf{X})$ o simplemente A. En la teoría de decisión, la anterior función se conoce como una regla de decisión.

Así como en la vida cotidiana, por la incertidumbre del futuro(en el ámbito estadístico, por la incertidumbre acerca del parámetro), toda acción que uno toma (toda estimación que uno provea) puede traer consigo un grado de falla o riesgo. Y es necesario tomar la acción óptima que de alguna forma minimice ese riesgo. Formalizando esta idea intuitiva, tenemos la función de pérdida L que asocia cada dupla de la acción tomada y el parámetro de interés θ, $(A,\ \theta)$ con un número no negativo que cuantifica la pérdida que ocasiona la acción (o la estimación) A con respecto al parámetro θ.

Es claro que se desea escoger aquella acción que minimice de alguna forma la pérdida que ésta ocasiona, pero la función L no se puede minimizar directamente, puesto que:

- En el ámbito de la estadística clásica, el parámetro θ se considera fijo, y los datos muestrales $\mathbf{X}$ aleatorios, así como la función de pérdida L depende de $\mathbf{X}$, entonces ésta también será una variable aleatoria, y no se puede minimizar directamente. Por lo tanto se define el riesgo o la pérdida promedio como la esperanza matemática de L; denotando el riesgo como R, éste está definido como $R = E(L)$ (la esperanza se toma con respecto

a la distribución probabilística de $\mathbf{X}$).

- En el ámbito de la estadística bayesiana, θ es una cantidad aleatoria, y la herramienta fundamental para conocer características de θ es su función de densidad posterior $p(\theta|\mathbf{X})$. En este caso, el riesgo R se define como

$$R = E(L) = \int L(A, \theta) p(\theta|\mathbf{X}) d\theta$$

En cualquier de los dos casos anteriores, buscaremos la estimación que minimice el riesgo R. Ilustramos los anteriores conceptos en los siguientes ejemplos tanto en la estadística clásica como en la estadística bayesiana.

Example 2.1. Sea X_i con $i = 1, \cdots, n$ una muestra aleatoria con media θ y varianza σ^2, ambas fijas, y suponga que se desea encontrar el mejor estimador de θ bajo la función de pérdida cuadrática dada por

$$L(A, \theta) = (A - \theta)^2$$

cuyo riesgo asociado está dado por $R = E(A-\theta)^2$. En primer lugar buscaremos dicho estimador dentro de todas las formas lineales de X_i, es decir, los estimadores de la forma $A = \sum_{i=1}^{n} c_i X_i$, de esta forma, el riesgo se puede expresar como

$$\begin{aligned} R = E(A-\theta)^2 &= Var(A) + (E(A) - \theta)^2 \\ &= \sum_{i=1}^{n} c_i^2 \sigma^2 + \theta^2 (\sum_{i=1}^{n} c_i - 1)^2 \end{aligned}$$

Y al buscar los coeficientes c_i que minimizan la anterior expresión, encontramos que $c_i = \theta^2/(\sigma^2 + n\theta^2)$ para todo i. Como estos coeficientes conducen a un estimador que depende del parámetro desconocido, concluimos que no hay ningún estimador que minimiza el riesgo.

Para encontrar una solución, es necesario restringir aún más el rango de estimadores, para eso, se restringe que $\sum_{i=1}^{n} c_i = 1$, de esta forma el riesgo está dado por $R = \sum c_i^2 \sigma^2$, y al minimizar $\sum c_i^2$ sujeto a la restricción de $\sum c_i = 1$. La solución está dada por $c_i = 1/n$ para todo i, y así encontramos que el mejor estimador (en el sentido de minimizar el riesgo de la función de pérdida cuadrática) dentro de todas formas lineales con $\sum c_i = 1$ es la media muestral $\bar{X}$.

Example 2.2. Suponga que se desea estimar un parámetro de interés θ en el contexto de la estadística bayesiana y denotamos la función de densidad posterior de θ como $p(\theta|\mathbf{X})$, entonces si utilizamos la función de pérdida cuadrática, entonces el riesgo asociado será

$$R = E(L(A, \theta)) = E(A - \theta)^2 = Var(\theta) + (E(\theta) - A)^2$$

que es minimizado si $A = E(\theta)$. Es decir la mejor acción para estimar θ es utilizar la esperanza de θ tomada con respecto a la distribución posterior $p(\theta|\mathbf{X})$.

Example 2.3. En el mismo contexto del ejemplo anterior, si cambiamos la función de pérdida a la siguiente

$$L(A,\theta) = |A-\theta| = (A-\theta)I_{(A\geq\theta)} + (\theta - A)I_{(\theta>A)}$$

Y el riesgo está dado por

$$\begin{aligned} R &= E(L(A,\theta)) \\ &= \int L(A,\theta)p(\theta|\mathbf{X})d\theta \\ &= \int_{(A\geq\theta)} (A-\theta)p(\theta|\mathbf{X})d\theta + \int_{(\theta>A)} (\theta - A)p(\theta|\mathbf{X})d\theta \end{aligned}$$

Derivando el riesgo con respecto a la acción A, se tiene que

$$\frac{\partial R}{\partial A} = \int_{(A\geq\theta)} p(\theta|\mathbf{X})d\theta - \int_{(\theta>A)} p(\theta|\mathbf{X})d\theta$$

Igualando a cero, tenemos que

$$\int_{(A\geq\theta)} p(\theta|\mathbf{X})d\theta = \int_{(\theta>A)} p(\theta|\mathbf{X})d\theta = 0.5$$

Y concluimos que la acción A que induce menor riesgo corresponde al percentil 50% o la mediana de la distribución posterior de θ.

De los anteriores ejemplos vemos que bajo un mismo contexto, cuando se utilizan diferentes funciones de pérdidas, también obtenemos distintas estimaciones.

Capítulo 3

Algunos resultados de probabilidad

A continuación se presentan definiciones y resultados de probabilidad en términos de notación se utilizará indistintamente la expresión de integral, $\int$, que implicará la integral, en el caso de las variables aleatorias continuas, o la sumatoria, en el caso de las variables aleatorias discretas.

Definition 3.1. Sean $\mathbf{X} = (X_1, \ldots, X_p)'$, $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ dos vectores aleatorios definidos sobre los espacios de muestreo $\mathcal{X}$, $\mathcal{Y}$, respectivamente. Suponga que la distribución conjunta de estos vectores aleatorios está dada por $p(\mathbf{X}, \mathbf{Y})$. La distribución marginal de $\mathbf{X}$ está dada por

$$p(\mathbf{X}) = \int p(\mathbf{X}, \mathbf{Y}) \, d\mathbf{Y} \tag{3.1}$$

y la distribución condicional de $\mathbf{X}$ dado $\mathbf{Y}$ como

$$p(\mathbf{X} \mid \mathbf{Y}) = \frac{p(\mathbf{X}, \mathbf{Y})}{p(\mathbf{Y})} \tag{3.2}$$

Proposition 3.1. *Suponga los vectores* $\mathbf{X}$, $\mathbf{Y}$ *y un tercer vector* $\mathbf{Z} = (Z_1, \ldots, Z_r)'$ *definido sobre el espacio de muestreo* $\mathcal{Z}$. *Entonces se tiene que*

$$p(\mathbf{X} \mid \mathbf{Z}) = \int p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \, d\mathbf{Y} \tag{3.3}$$

y

$$p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = \frac{p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})}{p(\mathbf{Y} \mid \mathbf{Z})} \tag{3.4}$$

Prueba. En primer lugar, nótese que

$$\begin{aligned}\int p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})\ d\mathbf{Y} &= \int \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Z})}\ d\mathbf{Y} \\ &= \frac{1}{p(\mathbf{Z})} \int p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\ d\mathbf{Y} \\ &= \frac{1}{p(\mathbf{Z})} p(\mathbf{X}, \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Z})\end{aligned}$$

Por otro lado,

$$\frac{p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})}{p(\mathbf{Y} \mid \mathbf{Z})} = \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Z})} \Big/ \frac{p(\mathbf{Y}, \mathbf{Z})}{p(\mathbf{Z})} = \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Y}, \mathbf{Z})} = p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z})$$

□

Definition 3.2. Sean $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ vectores aleatorios, se dice que $\mathbf{X}$ es condicionalmente independiente de $\mathbf{Y}$ con respecto a $\mathbf{Z}$ si satisfacen la siguiente expresión

$$p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Z}) p(\mathbf{Y} \mid \mathbf{Z}) \tag{3.5}$$

Proposition 3.2. *Si* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *con respecto a* $\mathbf{Z}$*, entonces se tiene que*

$$p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Z}) \tag{3.6}$$

Prueba. Como $p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = \dfrac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Z})}$, entonces

$$p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Y}, \mathbf{Z})} = \frac{p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) p(\mathbf{Z})}{p(\mathbf{Y}, \mathbf{Z})} = \frac{p(\mathbf{X} \mid \mathbf{Z}) p(\mathbf{Y} \mid \mathbf{Z})}{p(\mathbf{Y} \mid \mathbf{Z})} = p(\mathbf{X} \mid \mathbf{Z})$$

□

Proposition 3.3. *Si* $\mathbf{X}$ *es independiente de* $\mathbf{Y}$*, entonces* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *dada cualquier otro vector, digamos* $\mathbf{Z}$*.*

Prueba. Nótese que

$$p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) p(\mathbf{Y} \mid \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Z}) p(\mathbf{Y} \mid \mathbf{Z})$$

puesto que, utilizando la hipótesis de independencia, se tiene que

$$p(\mathbf{X} \mid \mathbf{Y}) = p(\mathbf{X})$$

□

3.1 Teorema de Bayes

Desde la revolución estadística de Pearson y Fisher, la inferencia estadística busca encontrar los valores que parametrizan a la distribución desconocida de los datos. El primer enfoque, propuesto por Pearson, afirmaba que si era posible observar a la variable de interés en todos y cada uno de los individuos de una población, entonces era posible calcular los parámetros de la distribución de la variable de interés; por otro lado, si sólo se tenía acceso a una muestra representativa, entonces era posible calcular una estimación de tales parámetros. Sin embargo, Fisher discrepó de tales argumentos, asumiendo que las observaciones están sujetas a un error de medición y por lo tanto, así se tuviese acceso a toda la población, es imposible calcular los parámetros de la distribución de la variable de interés.

Del planteamiento de Fisher resultaron una multitud de métodos estadísticos para la estimación de los parámetros poblacionales. Es decir, si la distribución de $\mathbf{Y}$ está parametrizada por $\theta = (\theta_1, \dots, \theta_K)$, $\theta \in \Theta$ con Θ el espacio paramétrico inducido por el comportamiento de la variable de interés, el objetivo de la teoría estadística inferencial es calcular una estimación $\hat{\theta}$ del parámetro θ por medio de los datos observados. En este enfoque, los parámetros se consideran cantidades fijas y constantes. Sin embargo, en la última mitad del siglo XX, algunos investigadores estadísticos comenzaron a reflexionar acerca de la naturaleza de θ y enfocaron la inferencia estadística de una manera distinta: asumiendo que la distribución de la variable de interés está condicionada a valores específicos de los parámetros. Es decir, en términos de notación, si la variable de interés es $\mathbf{Y}$, su distribución condicionada a los parámetros toma la siguiente forma $p(\mathbf{Y} \mid \theta)$. Esto implica claramente que en este nuevo enfoque la naturaleza de los parámetros no es constante sino estocástica.

En términos de inferencia para θ, es necesario encontrar la distribución de los parámetros condicionada a la observación de los datos. Para este fin, es necesario definir la distribución conjunta de la variable de interés con el vector de parámetros.

$$p(\theta, \mathbf{Y}) = p(\theta)p(\mathbf{Y} \mid \theta)$$

A la distribución $p(\theta)$ se le conoce con el nombre de distribución *previa* y en ella se enmarcan todas y cada una de las creencias que se tienen acerca del comportamiento estocástico del vector de parámetros antes de que ocurra la recolección de los datos y $p(\mathbf{Y} \mid \theta)$ es la distribución de muestreo o verosimilitud o distribución de los datos. Por otro lado, la distribución del vector de parámetros condicionada a los datos observados está dada por

$$p(\theta \mid \mathbf{Y}) = \frac{p(\theta, \mathbf{Y})}{p(\mathbf{Y})} = \frac{p(\theta)p(\mathbf{Y} \mid \theta)}{p(\mathbf{Y})} \tag{3.7}$$

A la distribución $p(\theta \mid \mathbf{Y})$ se le conoce con el nombre de distribución *posterior* y en ella se enmarcan las creencias actualizadas acerca del comportamiento