Estadística Bayesiana

Clase 1: Introducción a la Estadística Bayesiana

Isabel Cristina Ramírez Guevara

Escuela de Estadística
Universidad Nacional de Colombia, Sede Medellín

Medellín, 4 de agosto de 2020

Análisis de datos Bayesianos

El análisis de datos desde un enfoque Bayesiano trata de encontrar métodos prácticos para hacer inferencia utilizando modelos de probabilidad, tanto de las cantidades que se observan como de las que no se observan.

El proceso de análisis de datos utilizando un enfoque Bayesiano se puede dividir en los siguientes pasos:

a) Definir un modelo de probabilidad completo: una distribución de probabilidad conjunta para las cantidades observables y no observables en el problema.

Análisis de datos Bayesianos

El análisis de datos desde un enfoque Bayesiano trata de encontrar métodos prácticos para hacer inferencia utilizando modelos de probabilidad, tanto de las cantidades que se observan como de las que no se observan.

El proceso de análisis de datos utilizando un enfoque Bayesiano se puede dividir en los siguientes pasos:

a) Definir un modelo de probabilidad completo: una distribución de probabilidad conjunta para las cantidades observables y no observables en el problema.

b) Condicionar la distribución de las cantidades no observadas sobre un conjunto de cantidades observadas (datos). En otras palabras, encontrar la distribución posterior de las cantidades no observadas.

Análisis de datos Bayesianos

El análisis de datos desde un enfoque Bayesiano trata de encontrar métodos prácticos para hacer inferencia utilizando modelos de probabilidad, tanto de las cantidades que se observan como de las que no se observan.

El proceso de análisis de datos utilizando un enfoque Bayesiano se puede dividir en los siguientes pasos:

a) Definir un modelo de probabilidad completo: una distribución de probabilidad conjunta para las cantidades observables y no observables en el problema.

b) Condicionar la distribución de las cantidades no observadas sobre un conjunto de cantidades observadas (datos). En otras palabras, encontrar la distribución posterior de las cantidades no observadas.

c) Evaluar el ajuste del modelo y las implicaciones de la distribución posterior resultante.

Notación general

La inferencia estadística consiste en obtener conclusiones de cantidades desconocidas no observadas a partir de datos numéricos (muestra). En particular, se desea estimar:

i) Observaciones futuras del proceso.

ii) Parámetros.

Se va a utilizar la siguiente notación:

- θ: parámetro.
- $\mathbf{y}$: datos observados.
- $\tilde{y}$: predicciones.
- $p(\cdot)$: función de probabilidad o densidad de probabilidad.
- $p(\cdot|\cdot)$: probabilidad condicional.

Teorema de Bayes

Se tiene que $p(\theta, y)$: distribución conjunta de θ y y. De la definición de probabilidad condicional tenemos que:

$$p(y|\theta) = \frac{p(\theta, y)}{p(\theta)}$$

Teorema de Bayes

Se tiene que $p(\theta, y)$: distribución conjunta de θ y y. De la definición de probabilidad condicional tenemos que:

$$p(y|\theta) = \frac{p(\theta, y)}{p(\theta)}$$

$$p(\theta, y) = p(y|\theta)p(\theta)$$

Teorema de Bayes

Se tiene que $p(\theta, y)$: distribución conjunta de θ y y. De la definición de probabilidad condicional tenemos que:

$$p(y|\theta) = \frac{p(\theta, y)}{p(\theta)}$$

$$p(\theta, y) = p(y|\theta)p(\theta)$$

donde $p(\theta)$ es la distribución a priori de θ y $p(y|\theta)$ es la distribución de muestreo o verosimilitud.

Teorema de Bayes

Se tiene que $p(\theta, y)$: distribución conjunta de θ y y. De la definición de probabilidad condicional tenemos que:

$$p(y|\theta) = \frac{p(\theta, y)}{p(\theta)}$$

$$p(\theta, y) = p(y|\theta)p(\theta)$$

donde $p(\theta)$ es la distribución a priori de θ y $p(y|\theta)$ es la distribución de muestreo o verosimilitud.
El teorema de Bayes establece que la distribución posterior de θ condicionada sobre los valores observados y está dada por:

Teorema de Bayes

Se tiene que $p(\theta, y)$: distribución conjunta de θ y y. De la definición de probabilidad condicional tenemos que:

$$p(y|\theta) = \frac{p(\theta, y)}{p(\theta)}$$

$$p(\theta, y) = p(y|\theta)p(\theta)$$

donde $p(\theta)$ es la distribución a priori de θ y $p(y|\theta)$ es la distribución de muestreo o verosimilitud.
El teorema de Bayes establece que la distribución posterior de θ condicionada sobre los valores observados y está dada por:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

donde $p(y)= \sum_\theta p(\theta)p(y|\theta)$ si θ es discreto o $p(y)= \int_\theta p(\theta)p(y|\theta)$ si θ es contínuo.

Teorema de Bayes

Para un y fijo

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

donde $\propto$: proporcional a. El lado derecho de la ecuación corresponde a la distribución posterior no normalizada.

Al realizar inferencia con un enfoque Bayesiano inicialmente se deben definir $p(y|\theta)$ y $p(\theta)$ y luego realizar los cálculos necesarios para encontrar $p(\theta|y)$.

Teorema de Bayes

Para un y fijo

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

donde $\propto$: proporcional a. El lado derecho de la ecuación corresponde a la distribución posterior no normalizada.

Al realizar inferencia con un enfoque Bayesiano inicialmente se deben definir $p(y|\theta)$ y $p(\theta)$ y luego realizar los cálculos necesarios para encontrar $p(\theta|y)$.

Ejemplo

Se desea estimar la probabilidad, θ, de un evento, a partir del resultado de una sucesión de n ensayos Bernoulli. Por lo tanto $y_i = 1$ si se tiene un éxito o $y_i = 0$ si se obtiene un fracaso, con $i = 1, \cdots, n$. Encuentre la distribución posterior de θ si la distribución a priori es una U(0,1).

Principio de Verosimilitud

Al efectuar inferencias o tomar decisiones acerca de θ, después de que $\mathbf{y}$ ha sido observado, toda la información experimental relevante está contenida en la función de verosimilitud, evaluada en el vector observado $\mathbf{y}$. Además, dos funciones de verosimilitud contendrán la misma información sobre θ, en la medida que sean proporcionales.

Ejemplo

Pepe y Pacho sospechan que una moneda en particular está cargada y que θ la probabilidad de cara es mayor de 0.5. Ellos proponen dos experimentos diferentes para probar estadísticamente esta sospecha. Pepe va a lanzar la moneda 10 veces y su estadístico va a ser el número de caras en esos 10 intentos. Pacho lanzará la moneda hasta obtener 2 sellos y su estadístico va a ser el número de caras que obtenga en los lanzamientos.

Principio de Verosimilitud

Ejemplo

a) *Pepe obtuvo 8 caras en 10 lanzamientos. Obtenga el p-value de esta prueba.*

b) *En el momento en que Pacho obtuvo los dos sellos, había acumulado 8 caras. Encuentre el p-value de esta prueba.*

c) *¿Pepe y Pacho obtienen las mismas conclusiones? Suponga un nivel de significancia de 0.05.*

d) *Caro es bayesiana y decide asumir una distribución a priori no informativa para θ de:*

$$p(\theta) = 1 \quad 0 \leq \theta \leq 1$$

Muestre que la inferencia Bayesiana basada en la distribución posterior de θ va a ser exactamente la misma si Caro utiliza los resultados de Pacho o los de Pepe. Discuta esto como se relaciona con el principio de verosimilitud.

Predicción

Si y es desconocida pero observable, la distribución marginal de y o distribución predictiva a priori es:

$$p(y) = \int p(y, \theta) d\theta = \int p(y|\theta) p(\theta) d\theta$$

Después de que y ha sido observada, se puede predecir una cantidad desconocida $\tilde{y}$ observable, considerando la distribución predictiva posterior, que se encuentra sustituyendo en la ecuaciónn anterior a $p(\theta)$ por $p(\theta|y)$. La distribución predictiva posterior es:

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta$$

$$= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta$$

por independencia de $\tilde{y}$ y y se tiene

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

Ejemplo

La hemofilia es una enfermedad que muestra herencia recesiva ligada al cromosoma X, entonces un hombre que tiene el gen que causa la enfermedad en el cromosoma X es afectado, pero las mujeres que tienen el gen en uno de sus cromosomas X no es afectada. La enfermedad es fatal para la mujeres que heredan los dos genes y es un caso raro.

Suponga que una mujer tiene un hermano afectado por la enfermedad, por lo tanto la mujer tiene un 50 % de tener el gen "malo". Nos interesa el estado de la mujer. Los datos utilizados para determinar el estado de la mujer es la información de los dos hijos de la mujer los cuales están sanos.

i. *¿Cuál es la probabilidad posterior de que la mujer esté enferma?.*
ii. *La mujer tiene un tercer hijo sano, actualice la probabilidad encontrada en i.*