

VIDEO: INTRODUCCIÓN A LOS MODELOS BAYESIANOS USANDO PYMC3 - FERNANDO IRARRAZAVAL

Estadística bayesiana y el Teorema de Bayes

La estadística bayesiana resulta una rama de gran interés y de mucha utilidad en muchas áreas de estudio gracias a su capacidad para modelar y medir la incertidumbre de manera mas intuitiva y permitiendo responder preguntas mas prácticas, esto muestra una gran ventaja en comparación con la estadística frecuentista, pues, para responder las mismas preguntas, haría falta resolver cálculos muy complejos.

La estadística bayesiana se basa en el teorema de bayes, este teorema permite calcular la probabilidad de un sucesos, teniendo información de antemano sobre el mismo, la formula general es la siguiente:

$$P(\text{parámetro} \mid \text{datos}) = \frac{P(\text{datos} \mid \text{parámetro}) P(\text{parámetro})}{P(\text{datos})}$$

Esta fórmula contiene 3 componentes clave:

- **Probabilidad a posteriori** $P(\text{parámetro} \mid \text{datos})$: La información que se busca calcular, dado los datos observados.
- **Likelihood:** $P(\text{datos} \mid \text{parámetro})$: Modelo que describe como se generan o cual la probabilidad de ver los datos en distintos escenarios.
- **Priori:** $P(\text{datos})$: El conocimiento previo del parámetro.

Librería PyMC3

La librería PyMC3 es esta diseñada para construir modelos bayesiano de manera eficiente y muy flexible, pues permite definir modelos complejos y hacer inferencias sin la necesidad de resolver calculaos complicados (como en el caos del a estadística frecuentista).

Un ejemplos práctico del uso de la librería y del poder que tiene las herramientas que ofrece es la pandemia del COVID 19, donde se busca conocer la probabilidad de que un usuario de positivos a u test de COVID.

En este ejemplo, existen 3 posibles escenarios:

1. Modelo de prueba perfecto
2. Modelo con falsos positivos
3. Modelo con incertidumbre en la tasa de falsos positivos

Utilizar la estadística frecuentista para conocer la probabilidad p de que una prueba de positivo resultaría muy difícil, sin embargo, usando esta librería, eso no presenta ninguna dificultad.

1. Modelo de prueba perfecto

Este modelo considera que la prueba es completamente precisa, es decir, que no existen falsos positivos ni falsos negativos.

El pseudocódigo para crear el modelo es el siguiente:

1. *Importar pymc y Arviz*
2. *Definir datos*
3. *Definir la variable que se busca estimar y se le asigna una probabilidad a priori*
4. *Definir la distribución de los datos (Likelihood)*
5. *Aplicar pymc*
6. *Pymc devuelve miles de muestras de nuestros parámetros de interés*

El modelo calcula lo siguiente:

$$\text{Posterior de } p = \text{Prior}(\text{Conocimiento a priori}) \\ * \text{Likelihood} (\text{Modelo de como se generan los datos})$$

Este modelo permite responder la pregunta ¿Cuál es la probabilidad de que menos del 15% de la población esté infectada?

2. Modelo con falsos positivos

El segundo modelo toma en consideración la posibilidad de que existan los falsos positivos, en este caso, se estima que la tasa de falsos positivos es del 10%. En este caso, es necesario ajustar el modelo inicial (test perfecto) para calcular la probabilidad de que un usuario que hizo la prueba tenga COVID a pesar de la existencia de estos falsos positivos.

Nuevamente, es posible responder la pregunta ¿Cuál es la probabilidad de que menos del 15% de la población esté infectada?, además, se puede comparar como cambia dicho porcentaje en comparación con el primer modelo y brindar una respuesta más precisa o realista.

3. Modelo con incertidumbre en la tasa de falsos positivos

En el ultimo modelo, se introduce la incertidumbre al asumir que la tasa de falsos positivos no es fija (como en el modelo anterior), en este caso, se asume que sigue una distribución binomial y se ajusta el modelo. Para abordar dicha situación, basta con realizar pequeñas modificaciones al código de los modelos anteriores.

Estos modelos, al considerar la incertidumbre, permite realizar modelaciones mucho más robustas y realistas en situaciones en la que no se tiene certeza sobre algún parámetro importante, en el cual , los datos observados pueden variar mas de lo esperado o directamente se desconoce cuál es su comportamiento.

Conclusión

Finalmente, los modelos bayesiano ofrecen un enfoque poderos y muy flexible en situaciones bajo incertidumbre, situaciones en las cuales la estadística frecuentista recurriría a cálculos altamente complicados y poco prácticos para encontrar las respuestas que se buscan.

La librería PyMC3 ofrece herramientas que permiten realizar modelos y encontrar dichas soluciones sin la necesidad de manejar matemáticas muy complejas o avanzadas. Los ejemplos de las pruebas de COVID 19 ofrecen un ejemplo práctico y útil para observar la sencillez y eficacia de dicha librería (y del enfoque que ofrece la estadística bayesiana)