front_end_string = '''

Documento de introducción a la estadística bayesiana: explicación detallada e intuitiva con ejemplo discreto sencillo

1. Introducción

La estadística bayesiana ofrece un marco para **actualizar nuestras creencias** sobre parámetros desconocidos a medida que disponemos de nueva información. Se basa en el **Teorema de Bayes** y utiliza conceptos como:

- **Prior**: Creencia inicial sobre el valor de un parámetro antes de ver datos.
- **Likelihood (verosimilitud)**: Qué tan probable es observar los datos dados ciertos valores del parámetro.
- **Posterior**: Creencia actualizada sobre el valor del parámetro después de ver los datos.

El proceso bayesiano se entiende a menudo de manera iterativa: cada vez que obtenemos nuevos datos, **actualizamos** nuestra distribución de creencias (nuestra **posterior**) y esta posterior se convierte en la **prior** para el siguiente ciclo de actualización.

En este documento, nos centraremos en un ejemplo **discreto** y sencillo para poder hacer todos los cálculos a mano. Supondremos que el parámetro (por ejemplo, la probabilidad de que ocurra un evento) solo puede tomar unos cuantos valores discretos. En la práctica, muchos ejemplos se formulan con parámetros continuos (y se usan distribuciones conjugadas como Beta, etc.), pero para fines didácticos es más claro iniciar con un caso discreto paso a paso.

---

2. Repaso: El Teorema de Bayes

PROF

El Teorema de Bayes en su forma más común se expresa como:

$P(\theta \mid D) = \frac{P(D \mid \theta), P(\theta)}{P(D)},$

donde:

- θ es el parámetro desconocido (por ejemplo, la probabilidad de que un suceso ocurra).
- D son los datos observados.
- $P(\theta)$ es la **distribución a priori** (o simplemente el prior) de θ.
- $P(D \mid \theta)$ es la **verosimilitud** (o likelihood): la probabilidad de observar los datos D dado θ.
- $P(\theta \mid D)$ es la **distribución a posteriori** (o posterior) de θ, la probabilidad de θ después de observar D.
- $P(D)$ es la **probabilidad marginal** de los datos (o evidencia), que puede pensarse como un factor de normalización.

En la práctica, la fórmula más útil para la actualización bayesiana a menudo se escribe como:

$P(\theta \mid D) = \frac{P(D \mid \theta), P(\theta)}{\sum_{\theta'} P(D \mid \theta'), P(\theta')},$

cuando θ toma valores discretos (el denominador es la suma de todas las probabilidades de los datos bajo cada valor posible θ', lo que garantiza que la distribución posterior se normalice y sume a 1).

---

3. Ejemplo Discreto Sencillo

Supongamos que tenemos una moneda que **no sabemos** si está cargada o no (es decir, si la probabilidad de salir "cara" realmente es 0.5 o algún otro valor). Sin embargo, para simplificar el problema, **supondremos** que esta probabilidad θ solo puede tomar **5 valores posibles**:

$\theta \in {0.1,; 0.3,; 0.5,; 0.7,; 0.9}.$

Es decir, no consideramos ningún otro valor. Este es un caso sintético pero ilustrativo.

3.1. Definir la Prior

Antes de ver ninguna tirada de la moneda, expresamos nuestra **creencia inicial** de la forma más simple posible: tal vez consideramos que todos esos valores son igualmente probables. Entonces asignamos:

$P(\theta = 0.1) = 0.2, \quad
P(\theta = 0.3) = 0.2, \quad
P(\theta = 0.5) = 0.2, \quad
P(\theta = 0.7) = 0.2, \quad
P(\theta = 0.9) = 0.2.$

Es decir, cada valor tiene probabilidad 0.2. Esto es una **distribución a priori** (prior) uniforme sobre estos 5 puntos.

3.2. Observamos datos (Primer Bloque de Observaciones)

PROF

Supongamos que hacemos algunas tiradas y observamos la siguiente secuencia de resultados:

- Primer bloque de datos D_1: **H, T, T, H** (donde H = cara, T = cruz).

Contemos cuántas caras y cuántas cruces hay:

- Caras: 2
- Cruces: 2

3.2.1. Cálculo de la verosimilitud

Bajo la suposición de que cada tirada es independiente, si θ es la probabilidad de obtener cara en una tirada, entonces la probabilidad de observar la secuencia **H, T, T, H** (2 caras y 2 cruces en cualquier orden) es:

$P(D_1 \mid \theta) = \theta^{\text{(número de caras)}} \times (1-\theta)^{\text{(número de cruces)}}.$

En nuestro caso, hay 2 caras y 2 cruces, así que:

$P(D_1 \mid \theta) = \theta^2 \cdot (1-\theta)^2.$

3.2.2. Calcular la Posterior (no normalizada)

Para cada valor discreto de θ, calculamos:

$\text{Posterior no normalizada} = P(D_1 \mid \theta) \cdot P(\theta).$

Concretamente, usaremos la tabla siguiente para realizar todos los pasos:

θ	Prior $P(\theta)$	$\theta^2 (1-\theta)^2$	Producto = $P(D_1\mid \theta),P(\theta)$
0.1	0.2	$0.1^2 \cdot 0.9^2 = 0.1^2 \times 0.81 = 0.0081$	$0.2 \times 0.0081 = 0.00162$
0.3	0.2	$0.3^2 \cdot 0.7^2 = 0.09 \times 0.49 = 0.0441$	$0.2 \times 0.0441 = 0.00882$
0.5	0.2	$0.5^2 \cdot 0.5^2 = 0.25 \times 0.25 = 0.0625$	$0.2 \times 0.0625 = 0.0125$
0.7	0.2	$0.7^2 \cdot 0.3^2 = 0.49 \times 0.09 = 0.0441$	$0.2 \times 0.0441 = 0.00882$
0.9	0.2	$0.9^2 \cdot 0.1^2 = 0.81 \times 0.01 = 0.0081$	$0.2 \times 0.0081 = 0.00162$

Nota: Observa que $\theta=0.3$ y $\theta=0.7$ arrojan la misma verosimilitud, así como $\theta=0.1$ y $\theta=0.9$.

3.2.3. Normalización

PROF

La posterior sin normalizar para cada θ es la columna "Producto" en la tabla. Para tener una **distribución de probabilidad** válida, necesitamos que la suma de las probabilidades sea 1. Entonces:

1. Calculamos la **suma de todos los productos**:

   $\text{Suma} = 0.00162 + 0.00882 + 0.0125 + 0.00882 + 0.00162.$

   Hagamos la suma paso a paso:
   $0.00162 + 0.00882 = 0.01044,$
   $\quad 0.01044 + 0.0125 = 0.02294,$
   $\quad 0.02294 + 0.00882 = 0.03176,$
   $\quad 0.03176 + 0.00162 = 0.03338.$

   Aproximadamente $\text{Suma} = 0.03338$.

2. La **posterior** para cada θ se obtiene dividiendo cada valor entre esta suma:

$P(\theta \mid D_1) = \frac{P(D_1 \mid \theta), P(\theta)}{\text{Suma}}.$

Entonces, para cada θ:

θ	Posterior (no normalizada)	Posterior normalizada $\frac{\text{no norm}}{\text{Suma}}$
0.1	0.00162	$0.00162 / 0.03338 \approx 0.0485$
0.3	0.00882	$0.00882 / 0.03338 \approx 0.2642$
0.5	0.01250	$0.01250 / 0.03338 \approx 0.3743$
0.7	0.00882	$0.00882 / 0.03338 \approx 0.2642$
0.9	0.00162	$0.00162 / 0.03338 \approx 0.0485$

Verificamos que la suma de estas probabilidades sea 1 (o muy cerca de 1 por efectos de redondeo):
$0.0485 + 0.2642 + 0.3743 + 0.2642 + 0.0485 \approx 1.0.$

Por lo tanto, nuestra **distribución posterior** (o posteriori) después de observar **H, T, T, H** queda:

$\begin{aligned}
P(\theta=0.1 \mid D_1) &\approx 0.0485,\\
P(\theta=0.3 \mid D_1) &\approx 0.2642,\\
P(\theta=0.5 \mid D_1) &\approx 0.3743,\\
P(\theta=0.7 \mid D_1) &\approx 0.2642,\\
P(\theta=0.9 \mid D_1) &\approx 0.0485.
\end{aligned}$

3.2.4. Interpretación

Observamos que inicialmente todos los valores θ (0.1, 0.3, 0.5, 0.7, 0.9) eran igual de probables (0.2 cada uno). Pero tras ver la secuencia H, T, T, H, que da 2 caras y 2 cruces, la posterior "favorece" valores de θ cercanos a 0.5 (o sea, ni muy inclinados a cara ni muy inclinados a cruz). Por ello, $\theta=0.5$ obtiene la mayor probabilidad posterior. Los extremos 0.1 y 0.9 han bajado bastante, pero aún no son imposibles.

PROF

3.3. Segundo Bloque de Observaciones (Actualización Iterativa)

Supongamos ahora que continuamos el experimento y **hacemos 2 tiradas más**, observando:

- **H, H** (2 caras más).

Llamemos a estos nuevos datos D_2. Nuestra nueva muestra total de datos sería $D_1\cup D_2$ = (H, T, T, H) + (H, H). Sin embargo, podemos hacer la actualización en dos pasos:

1. Tomamos la **posterior** resultante de D_1 como la **prior** para el nuevo conjunto de datos D_2.
2. Calculamos la nueva posterior usando los mismos pasos.

3.3.1. Prior para la segunda actualización

La prior ahora es la **posterior** obtenida después de D_1:

$P(\theta=0.1) = 0.0485,\quad
P(\theta=0.3) = 0.2642,\quad
P(\theta=0.5) = 0.3743,\quad
P(\theta=0.7) = 0.2642,\quad
P(\theta=0.9) = 0.0485.$

3.3.2. Likelihood del nuevo bloque de datos D_2

Ahora, en D_2 tenemos **2 caras** (H, H). La probabilidad de observar este bloque **dado θ** es:

$P(D_2 \mid \theta) = \theta^2,$

porque son 2 caras seguidas y asumimos independencia.

3.3.3. Posterior no normalizada después de D_2

Calculamos para cada θ:

$\text{Posterior no norm}
= P(D_2 \mid \theta), P(\theta \mid D_1)
= \theta^2 \times \text{(Prior actual)}.$

Hagamos una tabla:

θ	Prior actual ($P(\theta\mid D_1)$)	θ^2	Producto = $\theta^2 \times P(\theta\mid D_1)$
0.1	0.0485	0.01	$0.01 \times 0.0485 = 0.000485$
0.3	0.2642	0.09	$0.09 \times 0.2642 = 0.023778$
0.5	0.3743	0.25	$0.25 \times 0.3743 = 0.093575$
0.7	0.2642	0.49	$0.49 \times 0.2642 = 0.129458$
0.9	0.0485	0.81	$0.81 \times 0.0485 = 0.039285$

PROF

3.3.4. Normalización

Sumamos todos los productos:

$0.000485 + 0.023778 + 0.093575 + 0.129458 + 0.039285.$

Hagamos la suma por partes:

- $0.000485 + 0.023778 = 0.024263$,
- $0.024263 + 0.093575 = 0.117838$,
- $0.117838 + 0.129458 = 0.247296$,

- $0.247296 + 0.039285 = 0.286581$.

Aproximadamente $\text{Suma} = 0.28658$.

La posterior normalizada es cada producto dividido entre 0.28658:

θ	Posterior no norm	Posterior final $P(\theta \mid D_1 \cup D_2)$ $\approx$
0.1	0.000485	$0.000485 / 0.28658 \approx 0.00169$
0.3	0.023778	$0.023778 / 0.28658 \approx 0.0830$
0.5	0.093575	$0.093575 / 0.28658 \approx 0.3266$
0.7	0.129458	$0.129458 / 0.28658 \approx 0.4520$
0.9	0.039285	$0.039285 / 0.28658 \approx 0.1370$

Verificamos la suma:
$0.00169 + 0.0830 + 0.3266 + 0.4520 + 0.1370 \approx 1.000.$

Entonces, **después de observar las dos tiradas adicionales (H, H)**, la distribución posterior es:

$\begin{aligned}
P(\theta=0.1 \mid D_1,D_2) &\approx 0.0017,\\
P(\theta=0.3 \mid D_1,D_2) &\approx 0.0830,\\
P(\theta=0.5 \mid D_1,D_2) &\approx 0.3266,\\
P(\theta=0.7 \mid D_1,D_2) &\approx 0.4520,\\
P(\theta=0.9 \mid D_1,D_2) &\approx 0.1370.
\end{aligned}$

3.3.5. Interpretación final

Después de ver un total de 4 caras y 2 cruces (en orden H, T, T, H, H, H), la probabilidad posterior se ha desplazado ahora **hacia valores más altos** de θ. Observamos que:

PROF

- $\theta=0.7$ se volvió el valor más probable (0.4520).
- $\theta=0.1$ prácticamente ha quedado descartado (solo 0.17%).
- $\theta=0.9$ tampoco es el más probable, pero tiene más peso que antes (13.7%); sin embargo, $\theta=0.7$ se ajusta más al total de 4 caras sobre 6 tiradas.

Si siguiéramos haciendo más tiradas, seguiríamos repitiendo este proceso: **la posterior resultante se convierte en la prior para la siguiente actualización** y así sucesivamente. Este es el ciclo fundamental de la inferencia bayesiana.

---

4. Conclusiones y Resumen del Proceso

1. **Elección del prior**: Arrancamos con una creencia inicial sobre θ. En el ejemplo, asignamos la misma probabilidad a 5 valores (prior uniforme).

2. **Cálculo de la verosimilitud (likelihood)**: Para los datos observados, calculamos la probabilidad de esos datos asumiendo cada valor de θ. En el ejemplo, se basó en el conteo de caras y cruces y la fórmula $\theta^{#\text{caras}} (1-\theta)^{#\text{cruces}}$.
3. **Actualización mediante el Teorema de Bayes**:
   $P(\theta \mid D) = \frac{P(D \mid \theta) , P(\theta)}{\sum_{\theta'} P(D \mid \theta') , P(\theta')}.$
   Esto nos dio una nueva distribución posterior que refleja cómo los datos observados modificaron nuestras creencias.
4. **Iteración**: La posterior se convierte en la prior para el siguiente bloque de datos. Observamos nuevos datos, repetimos el cálculo y obtenemos la siguiente posterior.

Ventajas principales:

- Permite **incorporar la incertidumbre inicial** (prior) de manera explícita.
- La actualización es un proceso **natural y acumulativo** (cada conjunto de datos ajusta un poco más la distribución).
- En este ejemplo discreto se ve claramente el "cambio" de la creencia inicial hasta la posterior que converge hacia los valores de θ más consistentes con los datos.

---

5. Reflexión Final

La estadística bayesiana no solo mira a la estimación puntual de un parámetro (e.g., "$\theta \approx 0.7$") sino a la **distribución** sobre el parámetro, reflejando la incertidumbre residual. Si hubiésemos recolectado muchísimos datos, la distribución posterior se concentraría más (sería más "pico" alrededor de un valor), reflejando que tenemos mayor confianza en que θ está cerca de ese valor.

El enfoque bayesiano es especialmente útil cuando:

- Queremos combinar **información previa** con datos nuevos.
- Tenemos pocos datos y deseamos una forma coherente de expresar incertidumbre.
- Necesitamos resultados probabilísticos completos y no solo estimaciones puntuales.

PROF

Este ejemplo sencillo ilustró el proceso en un caso discreto que podemos calcular fácilmente a mano. En problemas reales, donde θ puede variar continuamente, a menudo se usan familias de distribuciones conjugadas (Beta, Gamma, etc.) o métodos numéricos (MCMC, etc.) para realizar la actualización. Pero la **esencia conceptual** es la misma: **creencia previa** + **datos** → **creencia posterior**, todo regido por el Teorema de Bayes.

Referencias Breves

- **Teorema de Bayes**: $P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)}$.
- **Estadística Bayesiana**: Marco inferencial que se basa en la utilización de distribuciones de probabilidad tanto para los datos como para los parámetros.

¡Con esto, ya tienes un panorama claro de cómo funciona la inferencia bayesiana en su núcleo, con un ejemplo fácilmente calculable a mano!