Práctica 1

Regla de Bayes

El propósito de esta sección de la práctica es resolver situaciones que impliquen la aplicación de la Regla de Bayes como se presenta tradicionalmente en un curso de Probabilidad.

1. Demostración

Demuestra la validez de la siguiente expresión de la Regla de Bayes

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{k=1}^{K} P(A|B_k)P(B_k)}$$

donde A es un evento cualquiera y $\{B_1, \cdots, B_K\}$ forman una partición.

2. El test ~~in~~falible

En una población dada, una de cada mil personas tiene una enfermedad. Se toma una persona al azar de la población, se le aplica un test para detectar dicha enfermedad, y el resultado es positivo. El test se caracteriza por dar positivo el 99% de las veces que una persona tiene la enfermedad. Además, dicho test tiene una tasa de falsos positivos del 5%.

- ¿Cuál es la probabilidad de que la persona tenga efectivamente la enfermedad?
- Si realizamos el mismo análisis una segunda vez sobre el mismo paciente y obtenemos nuevamente positivo,
  - ¿Cuál seria la probabilidad que el paciente esté enfermo?
  - ¿Y si diera negativo?
  - ¿Es el *prior* el mismo cuando se analiza el resultado del segundo análisis que cuando solo se analiza el primero?

3. ¿Es verdad que existen los vampiros? Versión Crepúsculo

Edward quiere probarle a Bella que los vampiros existen. Según Bella, hay una probabilidad del 5% de que los vampiros existan. También cree que la probabilidad de que exista alguien con la piel brillante dado que los vampiros existen es del 70%, y que la probabilidad de que alguien tenga la piel brillante si los vampiros no existen es del 3%. Edward lleva a Bella al bosque y le muestra que de hecho su piel brilla como un ¿Cuál es la probabilidad que existan los vampiros?

4. Árboles enfermos

Un vivero de la ciudad se destaca por vender una variedad de árboles nativos, incluyendo al jacarandá, ceibo, ombú, entre otros. Lamentablemente, el 18% de los árboles del vivero estan infectados con moho. Los árboles enfermos se componen en un 15% por jacarandás, 80% de ceibos, y 5% de otras especies. Los árboles sanos se componen por un 20% de jacarandás, 10% de ceibos, y 70% de otras especies. Con el objetivo de monitorear cuanto se propagó la enfermedad, una de las personas que trabaja en el vivero selecciona al azar uno de los árboles para testear.

- ¿Cuál es la probabilidad *a priori* de que el árbol tenga moho?
- Resulta que el árbol seleccionado es un ceibo. ¿Cuál es la probabilidad de haber seleccionado un ceibo?

- ¿Cuál es la probabilidad *a posteriori* de que el ceibo seleccionado tenga moho?
- Compare las probabilidades *a priori* y *a posteriori* de que el árbol tenga moho. ¿Cómo afecta el análisis el saber que el árbol es un ceibo?

5. Transporte "El Impuntual"

Una cierta empresa de transporte regional, que decidimos llamar "El Impuntual", tiene servicios que van desde Rosario hasta Wheelwright varias veces al día, todos los días de la semana. Un 30% de los viajes salen a la mañana, otro 30% salen a la tarde, y el restante 40% salen a la noche. Los pasajeros suelen estar muy frustrados ya que un 25% de los viajes salen tarde. De estos viajes demorados, el 40% corresponden a la mañana, un 50% suceden a la tarde, y el 10% restante ocurre a la noche[1].

Lucio y Franco son dos amigos del pueblo, y se volvieron a sus casas en colectivos diferentes.

- Lucio se fue en uno de los colectivos de la mañana. ¿Cuál es la probabilidad que su viaje esté demorado?
- El colectivo de Franco no está demorado. ¿Cuál es la probabilidad de que esté viajando en uno de los colectivos de la mañana?

6. Bebé panda

Supongamos que hay dos especies de osos panda. Ambas especies son igual de frecuentes y viven en la misma región. Es más, lucen de la misma forma y comen la misma comida. Aún no existe una prueba genética que pueda diferenciarlos. Lo único que los diferencia es la cantidad de crías que suelen tener. Las madres de la especie A dan luz a mellizos el 10% del tiempo. Y las madres de la especie B dan a luz mellizos el 20% del tiempo. En todos los otros casos, estas madres dan a luz un solo bebé panda.

Usando un poco la imaginación, supongamos que somos la persona encargada de un programa de reproducción de pandas. Tenemos una panda femenina que acaba de dar a luz a un par de mellizos, pero no sabemos a que especie pertenece.

- ¿Cuál es la probabilidad que la mamá panda sea de la especie A?
- ¿Cuál es la probabilidad que vuelva a tener mellizos en la próxima parición?
- Un tiempo después sos encontramos con que en la segunda parición da a luz a un único bebé panda. ¿Cuál es la probabilidad de que este panda sea de la especie A?

7. Paraguas

Estás a punto de subir a un avión rumbo a Mendoza. Querés saber si tenés que llevar un paraguas o no. Llamás a tres amigos que viven en Mendoza y les preguntás si está lloviendo. Cada uno de ellos tiene una probabilidad de 2/3 de decirte la verdad y 1/3 de mentirte para hacerte una broma. Los tres responden que sí está lloviendo. ¿Cuál es la probabilidad de que realmente esté lloviendo en las Mendoza? Se puede asumir que en Mendoza llueve en 1 de cada 10 días.

> Ayuda
>
> Si LLL es "los tres amigos dijeron que llovía", buscamos $P(\text{lluvia} \mid LLL) = P(LLL \mid \text{lluvia})P(\text{lluvia})/P(LLL)$.
> $P(LLL \mid \text{lluvia})$ es la probabilidad de que ninguno de los tres mienta.
> $P(LLL) = P(LLL \mid \text{lluvia})P(\text{lluvia}) + P(LLL \mid \text{no lluvia})P(\text{no lluvia})$.

[1] Hay gente que dice que "cada dos por tres" te deja a pata. No nos vamos a pelear explicándoles que estan siendo demasiado exigentes, ya que un 25% también es un montón!

8. Sherlock

Dos personas dejaron rastros de sangre en la escena del crimen. La sangre de Guido, un sospechoso, es analizada y resulta ser de tipo '0'. Los rastros de sangre de la escena son de tipo '0' (un tipo común en la población, presente en el 60% de las personas) y de tipo 'AB' (un tipo raro, con una frecuencia del 1% en la población). ¿Estos datos representan evidencia de que Guido estaba presente en la escena del crimen?

> Ayuda
>
> Llamemos S a "Guido y otra persona estuvieron en la escena del crimen" versus S', "dos personas desconocidas estuvieron en la escena del crimen". Además, contamos con la evidencia E, "observar sangre 0 y sangre AB".
> Queremos saber qué es mayor, si $P(E \mid S)$ o $P(E \mid S')$ (si luego quisiéramos comparar $P(S \mid E)$ con $P(S' \mid E)$ necesitaríamos las probabilidades a priori de $P(S)$ y $P(S')$).
> Si $P(E \mid S) < P(E \mid S')$ entonces los datos aportan evidencia en contra de que Guido estaba en la escena del crimen (E es más probable bajo S' que bajo S).
> $P(E \mid S)$ es la probabilidad de observar sangre 0 y sangre AB, si Guido estaba en la escena del crimen. Esto es básicamente la probabilidad de encontrar sangre AB (porque la cuota de 0 ya está cubierta por Guido). Eso equivale a tener en la escena del crimen a una de las personas que tienen sangre AB (1/100 de la población). Luego, $P(E \mid S) = 0.01$.
> $P(E \mid S')$ es la probabilidad de observar sangre 0 y sangre AB, si Guido no estaba en la escena del crimen. Esto es la probabilidad de que la primera persona tuviera sangre 0 y la segunda AB más la probabilidad de que la primera persona tuviera sangre AB y la segunda sangre 0: $(60/100) \cdot (1/100) + (1/100) \cdot (60/100) = 0.012$.
> Para pensarlo intuitivamente: supongamos que hay 200 personas en un pueblo. 120 personas (Guido y 119 más) tienen sangre 0, 2 personas tienen sangre AB y el resto tiene otro tipo. Con Guido en la escena del crimen hay 2/200 personas que podrían haber estado con él; sin Guido en la escena, pudieron haber estado cualquiera de las 119 personas de sangre 0 con cualquiera de las 2 personas de sangre AB.

9. Hijos de la probabilidad

Nos encontramos con alguien en la calle y nos dice que tiene dos hijos. Le preguntamos si alguno de ellos es mujer y nos responde que sí. ¿Cuál es la probabilidad de que ambos sean niñas?

10. Los Reyes del Rock

Elvis Presley tenía un hermano varón que nació en el mismo parto pero que murió al poco tiempo. ¿Cuál es la probabilidad de que Elvis tuviera un gemelo? Alguna información adicional: en 1935, cuando Elvis nació, 1/3 de los hermanos del mismo parto eran gemelos y 2/3 mellizos; además, la probabilidad de que dos mellizos sean del mismo sexo biológico puede estimarse en 50%, mientras que dos gemelos son siempre del mismo sexo biológico.

11. ¿Alguien ordena las medias?

Dos cajones contienen medias. Uno de ellos tiene igual cantidad de medias blancas y negras. El otro contiene un número igual de medias rojas, verdes y azules. Se elige un cajón al azar, se sacan dos medias sin mirar y resultan ser las dos iguales. ¿Cuál es la probabilidad de que las medias sean blancas? Supóngase que sacar la primera media no altera las proporciones.

> **Ayuda**
>
> Sean los eventos
> C: elegir el cajón BN (y no el RVA)
> I: elegir un par de medias iguales
> B: elegir un par de medias blancas
> Matemáticamente, buscamos $P(B \mid I) = P(B, C \mid I) + P(B, C' \mid I)$ (básicamente estamos marginalizando la variable "cajón"). El segundo término es 0 porque B y C' no pueden darse nunca (nunca voy a sacar un par blanco del cajón RVA). Entonces queda $P(B \mid I) = P(B, C \mid I) = P(B \mid C, I)P(C \mid I)$, que son los dosvalores que encontramos arriba.
> ¿De dónde sale que $P(B, C \mid I) = P(B \mid C, I)P(C \mid I)$?
> $P(B, C \mid I) = P(B, C, I)/P(I)$ y el numerador, por regla de la cadena, es $P(A)P(C \mid I)P(B \mid C, I)$.

12. La Falacia del Fiscal

Sally Clark era una abogada británica que fue erróneamente sentenciada a prisión perpetua en 1999 por la muerte de sus dos hijos bebés. Su hijo mayor, Christopher, murió con 11 semanas en diciembre de 1996 y su hijo más joven, Harry, con 8 semanas en enero de 1998. Durante el juicio, la defensa argumentó que las muertes se debieron al síndrome de muerte súbita del lactante (SIDS). Clark fue condenada a partir del testimonio del pediatra Sir Roy Meadow, quien argumentó en la corte lo siguiente:

- En familias sanas, la chance de muerte por SIDS es de $\frac{1}{8500}$
- La probabilidad de dos muertes por SIDS en la misma familia es aproximadamente $\frac{1}{8500^2} \approx \frac{1}{73000000}$
- Es, por ende, muy poco probable que Clark sea inocente

Luego de pasar 3 años en prisión, Clark fue liberada en 2003 luego de que se determinara que el testimonio *experto* de Meadows era equivocado. Dos mujeres, a las cuales el testimonio de Meadows había enviado a prisión, también fueron liberadas.

i. Identifica una falla en la probabilidad de $\frac{1}{73000000}$ dada por Meadows.
ii. Incluso aceptando el número anterior como correcto, ¿cuál es el problema de interpretar esa probabilidad como la probabilidad de inocencia de Clark?

> **Ayuda**
>
> Aún asumiendo que $P(E \mid I)$ es lo que dice Meadows, esa no es la probabilidad de interés. La probabilidad que interesa es la de inocencia (o culpabilidad) dada la evidencia. Es decir, la probabilidad que nos interesa es $P(I \mid E)$ que la podemos escribir como $P(I \mid E) = P(E \mid I)P(I)/P(E)$.
> $P(I \mid E)$ es solo similar a $P(E\ midI)$ si $P(I)$ es similar a $P(E)$. Este no es el caso, ¿por qué?

Inferencia Bayesiana

En esta parte de la práctica, se le otorga un significado a las cantidades que aparecen en la Regla de Bayes modificando conceptualmente el enfoque de las situaciones problemáticas. Ahora los problemas se tratan de realizar inferencias sobre posibles causas de ciertos datos observados. Se incrementa el rigor matemático, aparecen distribuciones de probabilidad y la necesidad de dejar ciertos cálculos en manos de la computadora.

1. El lenguaje de las probabilidades

Escribir la expresión matemática para cada una de las siguientes descripciones verbales:

a. Probabilidad de un parámetro dados los datos observados.
b. La distribución de probabilidad de los parámetros antes de ver los datos.
c. La verosimilitud de los datos para un valor dado de los parámetros.
d. La probabilidad de una observación nueva luego de observar los datos.
e. La probabilidad de una observación antes de ver los datos.

2. Qué datazo me tiraste, rey

Los M&Ms azul fueron introducidos en el año 1995 (antes había dos tipos de marrón)

- Antes de 1995, la mezcla de colores en una bolsa de M&Ms era: 30% marron, 20% amarillo, 20% rojo, 10% verde, 10% naranja y 10% marrón *bronceado*.
- Luego de 1995, la mezcla pasó a ser: 24% azul, 20% verde, 16% naranja, 14% amarillo, 13% rojo y 13% marrón.

Un amigo tiene dos bolsas de M&M y nos dice que una bolsa es de 1994 y la otra es de 1996, pero no nos dice cuál es cuál. Nos da un M&M de cada bolsa: uno es amarillo y el otro es verde (ambos posiblemente estén vencidos). ¿Cuál es la probabilidad de que el amarillo venga de la bolsa de 1994?

3. La Gran Estafa

Hay dos monedas en una caja. Una de ellas es una moneda común y la otra es una moneda que tiene dos caras.

i. Se elige una moneda al azar, se arroja, y se obtiene cara. ¿Cuál es la probabilidad de que la moneda elegida sea la falsa?
ii. Se elige una moneda al azar y se arroja al aire tres veces, obteniéndose tres caras. ¿Cuál es la probabilidad de que la moneda elegida sea la falsa?

4. Vocabulario limitado

Supongamos que existe un idioma con seis palabras:

$$\{\text{perro, parra, farra, carro, corro, tarro}\}$$

Un análisis lingüístico exhaustivo de esta lengua ha descubierto que todas las palabras son igualmente probables, excepto por 'perro', que es α veces más probable que las otras.

Además:

- Cuando se tipean, un caracter se introduce erróneamente con probabilidad θ.
- Todas las letras tienen la misma probabilidad de producir un error de tipeo.
- Si una letra se tipeó mal, la probabilidad de cometer un error en otro caracter no cambia.
- Los errores son independientes a lo largo de una palabras.

i. ¿Cuál es la probabilidad de escribir correctamente 'tarro'?
ii. ¿Cuál es la probabilidad de tipear 'cerro' o 'curro' al querer escribir 'carro'?
iii. Utilizando la Regla de Bayes, desarrollar un corrector gramatical para esta lengua. Para las palabras tipeadas 'farra', 'birra' y 'locos', hallar la probabilidad de que cada palabra del diccionario sea la palabra que se había querido escribir. Utilizar las siguientes combinaciones de parámetros:

a. $\alpha = 2$ y $\theta = 0.1$

b. $\alpha = 50$ y $\theta = 0.1$
c. $\alpha = 2$ y $\theta = 0.9$

5. Que el árbol no tape el bosque

Sea $X_1 \sim \text{Bernoulli}(\theta)$ una variable que indica si una especie de árboles se halla en un determinado bosque y $\theta \in [0, 1]$ representa la probabilidad *a priori* de que la especie se encuentre en el bosque. Una investigadora selecciona una muestra de n árboles del bosque y encuentra que X_2 de ellas pertenecen a la especie de interés.

El modelo luego es

$$X_2|X_1 \sim \text{Binomial}(\lambda X_1, n) \quad \text{con } \lambda \in [0, 1]$$

λ representa la probabilidad de detectar la especie, dado que la especie se encuentra en el bosque.

Encuntre expresiones matemáticas en término de n, θ y λ para las siguientes probabilidades:

i. $P(X_1 = 0, X_2 = 0)$.
ii. $P(X_1 = 0)$.
iii. $P(X_2 = 0)$.
iv. $P(X_1 = 0|X_2 = 0)$.
v. $P(X_2 = 0|X_1 = 0)$.
vi. $P(X_1 = 0|X_2 = 1)$.
vii. $P(X_2 = 0|X_1 = 1)$.
viii. Explique de manera intuitiva cómo es que las probabilidades calculadas en (iv)-(vii) cambian según n, θ y λ.
ix. Asuma $\theta = 0.5$, $\lambda = 0.1$ y $X_2 = 0$ ¿Cuán grande debe ser n para que se puede concluir con 95% de confianza que la especie no se encuentra en el bosque?

6. ¡Ostras! ¡Estoy haciendo inferencia bayesiana!

En un estudio que utiliza métodos de la Estadística Bayesiana para predecir el número de especies que serán descubiertas en el futuro se reporta que la cantidad de especies marinas bivalvas[2] descubiertas cada año entre 2010 y 2015 fue 64, 13, 33, 18, 30 y 20.

Si se representa con Y_t a la cantidad de especies descubierta en el año t, y asumiendo:

$$\begin{aligned} Y_t|\lambda &\underset{iid}{\sim} \text{Poisson}(\lambda) \\ \lambda &\sim \text{Uniforme}(0, 100) \end{aligned}$$

Graficar la distribución *a posteriori* de λ.

7. Es negocio...

Sea n la cantidad desconocida de clientes que visitan una tienda en un dia cualquiera. El número de clientes que realizan una compra es Y y se cumple que

$$Y|\theta, n \sim \text{Binomial}(\theta, n)$$

donde θ es la probabilidad de compra, dado que se produce la visita a la tienda. La distribución *a priori* de n es $n \sim \text{Poisson}(5)$. Bajo el supuesto que θ es conocido y que n es desconocido, graficar la distribución *a posteriori* de n para todas las combinaciones de $Y \in \{0, 5, 10\}$ y $\theta \in \{0.2, 0.5\}$. Explique cual es del efecto de cambiar Y y θ sobre la distribución *a posteriori*.

[2] Una clase de molusco. El mejillón, la ostra y la almeja son bivalvos

8. Con amigos así, quién necesita enemigos

Un amigo arroja un dado y anota en secreto el número que sale (llamémoslo T). A continuación, nosotros, con los ojos vendados, arrojamos el dado varias veces. No podemos ver el número que sale pero nuestro amigo nos dice si el número que sacamos es mayor, menor o igual a T.

Supongamos que nos da la secuencia: G, G, C, I, C, C, C, I, G, C (siendo G más grande, C más chico e I igual). ¿Cuál es la distribución *a posteriori* de los valores de T?

> Ayuda
>
> Estamos tratando de hacer inferencias sobre T, es decir, la distribución a posteriori debe ser $P(T = 1), P(T = 2), \dots P(T = 6)$.
> Conviene analizar este problema en forma secuencial. Observamos G, ¿qué significa eso? ¿cómo obtenemos la verosimilitud? $P(G \mid T = 1) = 5/6$, $P(G \mid T = 2) = 4/6$, $P(G \mid T = 3) = 3/6$…

9. Orden en la sala

En las Jornadas Rosarinas de Ciencia de Datos, una expositora está dando una charla en un salón cuando el personal de seguridad la interrumpe porque cree que puede haber más de 1000 personas en la sala, superando el máximo permitido.

La expositora piensa que hay menos de 1000 personas y se ofrece a demostrarlo, aunque piensa que contarlas podría llevar mucho tiempo. Decide hacer un experimento:

- Pregunta cuántas personas nacieron el 11 de mayo. Dos personas levantan la mano.
- Pregunta cuántas personas nacieron el 23 de mayo. Una persona levanta la mano.
- Pregunta cuántas personas nacieron el 1 de agosto. Nadie levanta la mano.

¿Cuántas personas hay en la sala? O, mejor dicho, ¿cuál es la probabilidad de que haya más de 1000 personas en la sala?

> Ayuda
>
> Estamos tratando de hacer inferencias sobre la cantidad de personas X (que puede ser un número entre, digamos, 1 y 3000).
> Conviene analizar el problema en forma secuencial. Para la primera observación (dos personas cumpliendo años el mismo día), la verosimilitud es la probabilidad de que haya dos personas con el mismo cumpleaños si en la sala hay 1, 2, 3, … personas. Lo podemos pensar como una binomial
> Si $X = 1$, $P(Y = 2 \mid X = 1) = 0$
> Si $X = 2$, $P(Y = 2 \mid X = 2) =$ `dbinom(2, size = 2, prob = 1/365)`
> Si $X = 100$, $P(Y = 2 \mid X = 100) =$ `dbinom(2, size = 100, prob = 1/365)`

10. *House of Cards*

Hay 538 miembros en el Congreso de Estados Unidos. Supongamos que se auditan sus inversiones y se encuentra que 312 de ellos obtuvieron rendimientos por encima del mercado. Asumamos que un miembro honesto del Congreso tiene solo una probabilidad del 50% de tener rendimientos por encima del mercado, pero uno deshonesto que opera con información confidencial tiene una chance del 90% de hacerlo. ¿Cuántos miembros del Congreso son honestos?

11. Puede fallar...

Cansada de los experimentos de arrojar una moneda cientos de veces al aire, una estudiante diseña un sistema de reconocimiento de imágenes que determina si salió cara o ceca y registra el resultado.

Lógicamente, el sistema diseñado no es perfecto sino que presenta una tasa de error. En particular, la probabilidad de que clasificar mal es de 0.2 (20% de las veces que sale cara, el sistema dice ceca, y viceversa).

Se arroja la moneda 250 veces y el sistema detecta 140 caras,

i. ¿Cuál es la distribución *a posteriori* de θ, la probabilidad de obtener cara?
ii. ¿Qué ocurre a medida que la probabilidad de clasificar mal varía?

12. ¡Saludos a los cubos con puntos! (...) Serán dados

Dos dados de seis caras son arrojados. Se sabe que la suma de los dos puntajes obtenidos es 9. ¿Cuál es la distribución *a posteriori* de los puntajes de los dados?

> Ayuda
>
> Estamos tratando de hacer inferencias sobre D_1 y D_2, los valores de cada uno de los datos. Definir la distribución a priori no es complicado (es una distribución sobre D_1 y D_2).
> El truco está en determinar la verosimilitud de cada par (D_1, D_2) luego de haber observado un 9.
> ¿Cuál es la probabilidad de observar 9 en la suma de los dos dado que ...?
> $D_1 = 1$ y $D_2 = 1-$ la probabilidad de observar 9 es 0
> $D_1 = 1$ y $D_2 = 2-$ la probabilidad de observar 9 es 0
> $D_1 = 4$ y $D_2 = 5-$ la probabilidad de observar 9 es 1
> $D_1 = 5$ y $D_2 = 4-$ la probabilidad de observar 9 es 1

Conceptuales

En esta sección, se nos invita a pensar sobre las características de la Estadística Bayesiana. En lugar de encontrar una respuesta única mediante cálculos matemáticos, se necesita comprender en profundidad tanto el enfoque frecuentista como el bayesiano para interpretar estas visiones en diferentes escenarios.

1. Voy a conseguir esa pasantía

La empresa de tecnología en la que todo el mundo quiere trabajar tiene varias vacantes para pasantes en ciencia de datos. Luego de leer la descripción de la búsqueda, te das cuenta que sos una persona calificada para el puesto: estos son tus **datos**. Tu objetivo es averiguar si te van a ofrecer el puesto: esta es tu **hipótesis**.

i. Desde la perspectiva de una persona con un razonamiento frecuentista, ¿Qué es lo que se responde al evaluar la hipótesis de que te ofrecen el puesto?
ii. Repita el punto anterior considerando la perspectiva de una persona con un razonamiento Bayesiano.
iii. ¿Qué pregunta tiene más sentido responder: la frecuentista o la Bayesiana? Justifica tu respuesta.

2. Beneficios de la Estadística Bayesiana

Una amiga te cuenta que está interesada en aprender más sobre Estadística Bayesiana. Explícale lo siguiente:

i. ¿Por qué es útil el enfoque Bayesiano?
ii. ¿Cuáles son las similitudes entre el enfoque frecuentista y el Bayesiano?