Práctica 1.
Especificación y evaluación de argumentos causales.

Docente: Gustavo Landfried

Inferencia Bayesiana Causal 1
1er cuatrimestre 2025
UNSAM

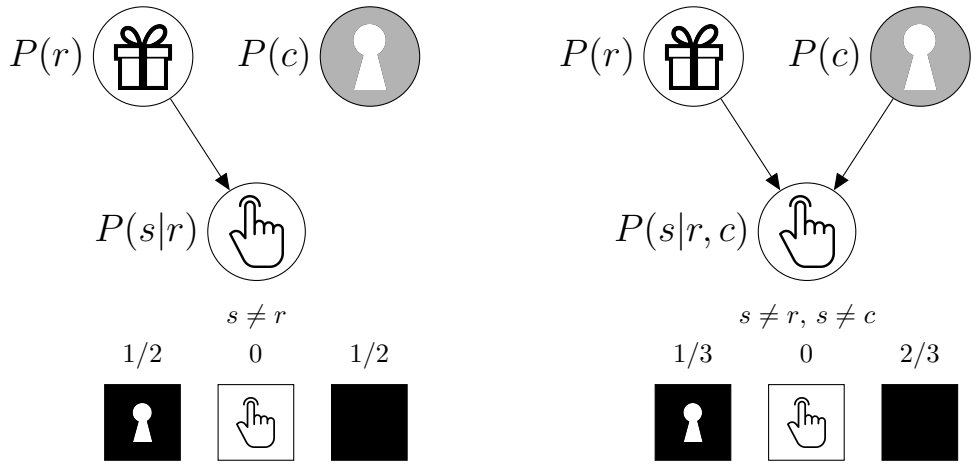
Índice

1. Modelo Base vs Modelo Monty Hall **2**
1.1. Definir la distribución de creencia conjunta como producto de las distribuciones condicionales del modelo . . . 3
1.2. Simular datos con el modelo Monty Hall . . . 3
1.3. Calcular la predicción a priori que hace cada uno de los modelos sobre la totalidad de la base de datos simulada . . . 4
1.4. Calcular la predicción de los datos con la contribución de todos los modelos. . . . 4
1.5. Calcular y graficar el posterior de los modelos . . . 5

2. Modelo Alternativo **5**
2.1. Calcular el posterior sobre la memoria p. . . . 7
2.2. Calcular la predicción de un episodio dado los datos de los episodios anteriores . . 8
2.3. Calcular la predicción que hace el modelo alternativo M_A sobre todo el conjunto de datos. . . . 8
2.4. Comparar el desempeño del modelo alternativo respecto de los modelos Base y el modelo MontyHall. . . . 9
2.5. Calcular la predicción típica que hace el modelo de los episodios. . . . 9
2.6. Calcular el posterior en los primeros episodios y graficar . . . 10

1. Modelo Base vs Modelo Monty Hall

Los datos no hablan por si solos. En todas las ciencias con datos se proponen teorías causales mediantes las cuales se interpretan los datos. En la siguiente figura se puede observar la especificación gráfica del modelo "Monty Hall" (derecha) y el modelo "Base" (izquierda) vistos en la primera semana mediante la notación de redes bayesianas. Abajo de ellos se muestra la distribución de creencias *a posteriori* sobre la posición del regalo luego de haber reservado la caja 1 y luego de que nos hayan mostrado que en la caja 2 no estaba el regalo, $P(r|s = 2, c = 1)$.



(a) Modelo Base (M_0)

(b) Modelo Monty Hall (M_1)

Figura 1: Modelos causales probabilísticos alternativos representados mediante la notación de redes bayesianas causales.

Las redes bayesianas son un método gráfico de especificación matemática de la distribución de probabilidad conjunta entre las variables (nodos de la red) mediante la definición de una distribución de probabilidad condicional por cada una de las variables, donde las flechas representan las dependencias condicionales entre variables. Una misma distribución conjunta se puede descomponer de muchas formas alternativas. Una red bayesiana es causal solo cuando la descomposición se corresponde con la semántica causal, es decir, las distribuciones de probabilidad condicional representan mecanismos causales probabilísticos entre causas y efectos. Los modelos Base y Monty Hall son ejemplos de redes bayesianas causales.

Las redes bayesianas causales encapsulan todas las hipótesis de investigación que se utilizan para interpretar los datos y sacar conclusiones sobre las hipótesis ocultas. La única restricción que supone el modelo Base (izquierda) es que la pista s no puede señalar la caja en la que se encuentra el regalo $s \neq r$. El modelo Monty Hall (derecha) incluye esta restricción y le agrega una restricción adicional, que la pista s tampoco puede señalar la caja que hemos reservado previamente $s \neq c$.

Sin embargo, los modelos causales también son hipótesis (que contienen hipótesis en su interior, las variables ocultas) y una de las tareas más importantes de todas las ciencias con datos es evaluar los modelos causales alternativos. El objetivo de esta guía es actualizar nuestras creencias sobre los modelos causales alternativos luego de observar un conjunto datos, $P(\text{Modelo}|\text{Datos})$.

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Datos}|\text{Modelo})P(\text{Modelo})}{P(\text{Datos})}$$

Para ello deberemos calcular:

- La predicción que hace el modelo sobre los datos: $P(\text{Datos}|\text{Modelo})$
- La predicción de los datos realizada con la contribución de todos los modelos: $P(\text{Datos})$
- La creencia previa "honesta" sobre los modelos: $P(\text{Modelo})$

1.1. Definir la distribución de creencia conjunta como producto de las distribuciones condicionales del modelo

Las redes bayesianas causales de la figura 1 representan la especificación matemática de dos argumentos causales alternativos. Las distribución de probabilidad conjunta de los modelos, $P(r, c, s|M)$, se pueden reconstruir como el producto de las distribuciones de sus mecanismos causales probabilísticos.

$$\underbrace{P(r,c,s|M_0)}_{\substack{\text{Prior conjunto}\\ \text{hipótesis en } M_0}} = \underbrace{P(r|M_0)P(c|M_0)P(s|r,M_0)}_{\substack{\text{Relaciones causales probabilísticas}\\ \text{propuestas por el modelo } M_0}} \quad , \quad \underbrace{P(r,c,s|M_1)}_{\substack{\text{Prior conjunto}\\ \text{hipótesis en } M_1}} = \underbrace{P(r|M_1)P(c|M_1)P(s|r,c,M_1)}_{\substack{\text{Relaciones causales probabilísticas}\\ \text{propuestas por el modelo } M_1}}$$

Ambos modelos suponen que r y c son variables independientes. Maximizando incertidumbre entre las 3 opciones obtenemos distribuciones condicionales a priori sobre r y c iguales en ambos modelos.

$$P(r|M) = P(r) = \begin{array}{c|c|c} r=0 & r=1 & r=2 \\ \hline 1/3 & 1/3 & 1/3 \end{array} \qquad P(c|M) = P(c) = \begin{array}{c|c|c} c=0 & c=1 & c=2 \\ \hline 1/3 & 1/3 & 1/3 \end{array}$$

La única diferencia entre los modelos aparece en la distribución condicional sobre la pista. El modelo Base M_0 supone que s depende únicamente de r, $(s \neq r)$ mientras que el modelo Monty Hall M_1 considera que s depende tanto de r como de s, $(s \neq r,\ s \neq c)$. Maximizando incertidumbre entre las opciones disponibles obtenemos la siguiente distribución condicional para el modelo Base (que solo depende del regalo r).

$$P(s|r, M_0) = \begin{array}{c|ccc} & s=0 & s=1 & s=2 \\ \hline r=0 & 0 & 1/2 & 1/2 \\ \hline r=1 & 1/2 & 0 & 1/2 \\ \hline r=2 & 1/2 & 1/2 & 0 \end{array}$$

Notar que hay una distribución de probabilidad por cada uno de los condicionales (cada regalo), lo que implica que (los renglones) tiene que integrar 1.

En el modelo Monty Hall el condicional depende de dos variables, el regalo r y la caja elegida c. Para simplificar, mostraremos los valores cuando $c = 1$.

$$P(s|r, c=1, M_1) = \begin{array}{c|ccc} (c=0) & s=0 & s=1 & s=2 \\ \hline r=0 & 0 & 1/2 & 1/2 \\ \hline r=1 & 0 & 0 & 1 \\ \hline r=2 & 0 & 1 & 0 \end{array}$$

Nuevamente, notar que cada renglón suma 1 pues cada condicional representa una distribución de probabilidad distinta.

1.2. Simular datos con el modelo Monty Hall

Antes de evaluar los modelos necesitamos un conjunto de datos que provengan de la realidad causal subyacente. Podríamos buscar los datos reales del programa de televisión Monty Hall y revisar si efectivamente el modelo Monty Hall propuesto es mejor que el modelo Base. Aquí vamos a suponer que nuestro modelo Monty Hall representa perfectamente la realidad causal subyacente y vamos a generar los datos a partir de él.

Las redes bayesianas causales son siempre modelos generativos a partir de los cuales se pueden simular datos. *Ancestral sampling* es el proceso que imita la generación de datos en el mundo real siguiendo el orden causal del grafo. Primero se muestrean los valores de los nodos raíz, las variables que no dependen de ninguna otra. Y luego se van muestreando las variables para las cuales ya tienen definidas el valor de todas sus causas. Cuando todos los nodos tengan un valor, se obtiene una única muestra completa y consistente con las relaciones causales y probabilísticas definidas por la red.

Generar un conjunto de datos con $T = 16$ episodios.

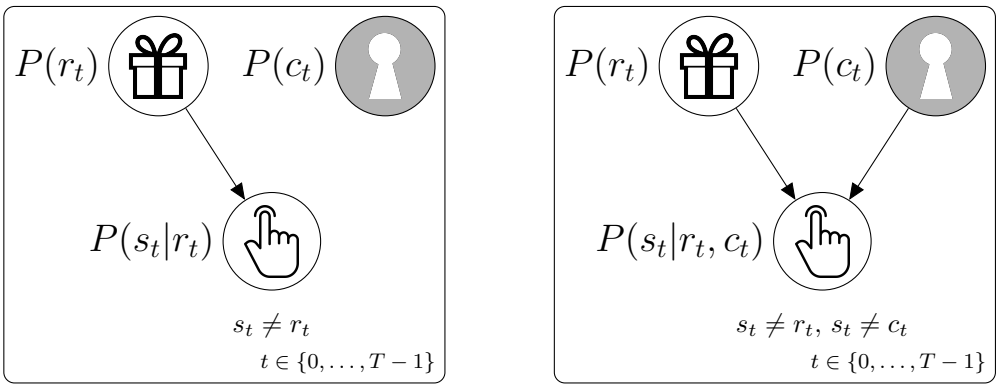
$$\text{Datos} = \{(c_0, s_0, r_0), \ldots, (c_{T-1}, s_{T-1}, r_{T-1})\}$$

1.3. Calcular la predicción a priori que hace cada uno de los modelos sobre la totalidad de la base de datos simulada

Ahora sí, podemos calcular la predicción del conjunto de datos que hace cada uno de los modelos con la contribución de todas sus hipótesis internas.

$$P(\text{Datos} = \{\underbrace{(c_0, s_0, r_0)}_{\substack{\text{Primer}\\ \text{episodio}}}, \underbrace{(c_1, s_1, r_1)}_{\substack{\text{Segundo}\\ \text{episodio}}}, \ldots \underbrace{(c_{T-1}, s_{T-1}, r_{T-1})}_{\substack{\text{T-ésimo}\\ \text{episodio}}}\}|\text{Modelo})$$

Los modelos causales expresados en la figura 1 proponen relaciones causales probabilísticas entre las variables al interior de un episodio. En principio, estos modelos sólo están definidos para un único episodio. Para extenderlos a T episodios vamos a considerar que contamos con T repeticiones de esa misma estructura causales. Las repeticiones se especifican gráficamente mediante el uso de "placas".



(a) Modelo Base (M_0)

(b) Modelo Monty Hall (M_1)

Figura 2: Extensión de los modelos causales alternativos a T episodios mediante la notación de "placas". El subíndice t representa las repeticiones.

Dado que entre episodios no hay flechas que vinculen entre sí las estructuras causales, lo que ocurra en un episodio no va a influir en lo que ocurra en otro episodio (los episodios son independientes entre sí). Esto permite descomponer la predicción sobre el conjunto de datos sobre todos los episodios como el producto de las predicciones que lo modelos hacen al interior de cada episodio.

$$P(\text{Datos}|\text{Modelo}) = \prod_{t \in \{0,\ldots,T-1\}} P(c_t|\text{Modelo}) P(s_t|c_t, \text{Modelo}) P(r_t|s_t, c_t, \text{Modelo})$$

Calcular la evidencia $P(\text{Datos}|\text{Modelo})$. Guardar el valor de la evidencia a medida que vamos agregando datos en la secuencia de predicciones (por evento).

1.4. Calcular la predicción de los datos con la contribución de todos los modelos.

Para actualizar la creencia de los modelos vamos a necesitar la probabilidad de los datos, $P(\text{Datos})$, que no es más que la predicción hecha con la contribución de todos los modelos.

$$P(\text{Datos}) \overset{\substack{\text{Regla de}\\ \text{la suma}}}{=} \sum_{\text{Modelo}} P(\text{Modelo}, \text{Datos}) \overset{\substack{\text{Regla de}\\ \text{la producto}}}{=} \sum_{\text{Modelo}} \underbrace{P(\text{Datos}|\text{Modelo})}_{\substack{\text{Predicción hecha}\\ \text{por el modelo}}} \underbrace{P(\text{Modelo})}_{\substack{\text{Creencia en}\\ \text{el modelo}}}$$

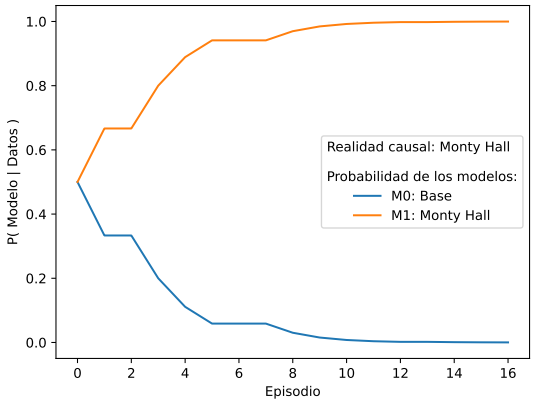
Aprovechar la secuencia de valores de la evidencia (predicciones hechas por el modelo en cada uno de los episodios) para calcular cómo se va actualizando el ensamble de predicciones $P(\text{Datos})$ a medida que se van incorporando nuevos episodios.

1.5. Calcular y graficar el posterior de los modelos

Ahora sí. Tenemos todos los elementos necesarios para calcular el posterior de los modelos.

$$P(\text{Modelo}|\text{Datos}) = \frac{P(\text{Modelo}, \text{Datos})}{P(\text{Datos})}$$

Para graficar cómo se va actualizando el posterior deberemos tener guardado el valor del posterior luego de observar cada uno de los episodios.



2. Modelo Alternativo

Se provee una archivo de datos `NoMontyHall.csv` que contienen 2000 episodios. Los datos fueron generados con la siguiente realidad causal subyacente. La persona que da la pista tiene una probabilidad $p \in [0, 1]$ de acordarse de tener en cuenta la caja que reservamos a la hora de dar la pista. Esta probabilidad es general a todos los episodios. En cada episodio particular, la persona se acuerda o no de tener en cuenta la pista, $a \in \{0, 1\}$. Cuando se acuerda, la persona usa la distribución de probabilidad condicional del modelo Monty Hall para dar la pista. Cuando se olvida usa la distribución de probabilidad condicional del modelo Base para dar la pista.

Para especificar matemáticamente el modelo causal vamos a usar la notación gráfica que se conoce como *factor graph*. Los factor graph, a diferencia de las redes bayesianas, incorporan los mecanismos causales probabilísticos (sus distribuciones de probabilidad condicional) como nodos de la red causal, formando un grafo bipartito en el cual las variables quedan vinculadas con las distribuciones de probabilidad de las cuales son parámetro. Al igual que las redes bayesianas, en los factor graph el producto de las distribuciones de probabilidad condicional es la especificación matemática de la distribución de probabilidad conjunta. Sin embargo, los factor graph tienen varias ventajas respecto de la notación de redes bayesianas. En particular, esta notación permite definir mecanismos causales dinámicos, cambian en función del contexto como son las intervenciones externas sobre mecanismos causales específicos. Esto se puede especificar mediante la notación de compuertas introducida en el artículo *Causality with gates*.

En el modelo alternativo, ahora la pista depende de 3 variables, por lo que su distribución de probabilidad condicional tiene 3 variables en el condicional. Sin embargo, podemos ser más específicos y mostrar explícitamente que esa distribución de probabilidad condicional es en realidad una mezcla de mecanismos causales probabilísticos.

$$P(s_t|r_t, c_t, a_t) = P(s_t|r_t)^{\mathbb{I}(a_t=0)} P(s_t|r_t, c_t)^{\mathbb{I}(a_t=1)} \tag{1}$$

Si la persona se olvida de tener en cuenta la caja elegida, $a_t = 0$, la persona de la pista siguiendo el mecanismo causal del modelo Base, $P(s_t|r_t)$. Si la persona se acuerda de tener en cuenta la caja elegida, $a_t = 1$, la persona de la pista siguiendo el mecanismo causal del modelo Monty Hall, $P(s_t|r_t, c_t)$.

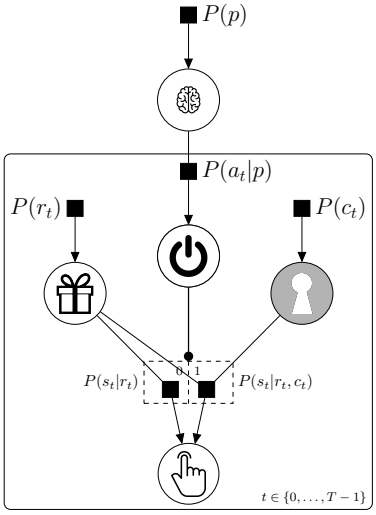


Figura 4: Realidad causal subyacente que generó el conjunto de datos del archivo `NoMontyHall.csv`, especificado mediante la notación de *factor graph*.

Luego, la probabilidad conjunta es

$$P(r_0, c_0, s_0, a_0, \ldots, p|M_2) = P(p) \prod_{t=0}^{T} P(r_t)P(c_t)P(s_t|r_t, M_0)^{1-a_t} P(s_t|r_t, c_t, M_1)^{a_t} P(a_t|p) \quad (2)$$

A diferencia de lo que ocurre en los modelos Base y Monty Hall, en el cual los datos de los diferentes episodios son independientes entre sí, en este modelo hay una variable, la probabilidad p de acordarse, que es común a todos los episodios y los conectan entre sí. Si abrimos las placas con el subíndice t que representa la repetición de los episodios vamos a ver que ellos quedan conectados entre sí por la variable p.

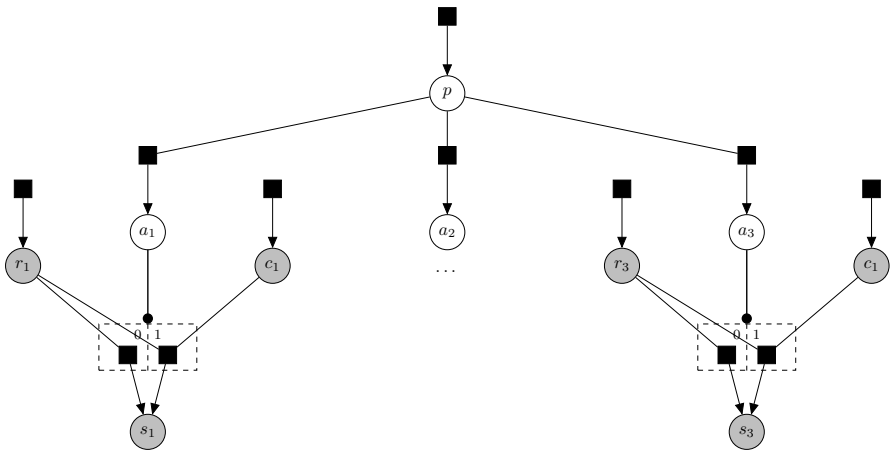


Figura 5: Factor graph del modelo alternativo desplegado. Las variables en blanco son ocultas, y las variables en gris son observadas (disponibles en la base de datos).

El archivo `NoMontyHall.csv` tiene los datos de los episodios: la caja que reservamos c_t, la pista que nos ofrecen s_t, y la posición del regalo r_t. Las variables que hemos agregado para modelar el problema, la probabilidad de acordarse p y las variables de cada episodio que representan si que efectivamente se acuerda o no a, permanecen ocultas. Las variables que se grafican en blanco representan variables ocultas (como la probabilidad de acordarse p y las variables que representan si que efectivamente se acuerda o no a) y las variablesque se grafican en fgris representa variables observadas (el regalo, la caja reservada y la pista).

2.1. Calcular el posterior sobre la memoria p.

A diferencia de los modelos Base y Monty Hall, en el modelo alternativo los episodios están conectados entre sí por la memoria p. Cada vez que recibimos nuevos datos, deberemos actualizar la distribución de creencias sobre la memoria p, y usar ese posterior como prior del siguiente evento.

¿Cómo podemos calcular el posterior de la memoria p? Recordar que el sistema de razonamiento para contextos de incertidumbre tiene solo dos reglas: la regla del producto (distribución de probabilidad condicional) mediante la cual preservamos la creencia previa que sigue siendo compatible con el datos; y la regla de la suma (distribución de probabilidad marginal) mediante la cual predecimos eventos aun no observados mediate la contribución de todas las hipótesis del modelo. Si tenemos la distribución de probabilidad conjunta estas dos reglas nos permiten derivar todas cualquier tipo de conclusión. Revisemos entonces la distribución de probabilidad conjunta del modelo alternativo que se presentó en la ecuación 2, que aquí volvemos a escribir de forma levemente simplificada.

$$P(r_0, c_0, s_0, a_0, \ldots, p) = P(p) \prod_{t=0}^{T} \underbrace{P(r_t)P(c_t)P(s_t|r_t, c_t, a_t)P(a_t|p)}_{P(r_t, c_t, s_t, a_t|p)} \tag{3}$$

Recordar que el posterior sobre hipótesis oculta dado los datos no es más que la distribución de probabilidad condicional, que tiene en el numerador la probabilidad conjunta entre la hipótesis y los datos, y tiene en el denominador la probabilidad conjunta de los datos (integrando todas las posibles hipótesis). Aquí los datos contienen solo la posición del regalo r_t, la caja que reservamos c_t y la pista que nos dan s_t. Nunca sabemos si la persona se acordó o no de tener en cuenta la caja que reservamos para dar la pista, por lo que a_t permanece como variable oculta. Luego, por las reglas de la probabilidad, podemos calcular la conjunta entre la hipótesis p y los eventos observados r_t, c_t, s_t integrando (regla de la suma) las hipótesis ocultas a_t.

$$P(p|\underbrace{r_0, c_0, s_0, \ldots}_{\text{Datos}}) = \frac{\overbrace{\sum_{a_0 \ldots a_T} P(r_0, c_0, s_0, a_0, \ldots, p)}^{P(r_0, c_0, s_0, \ldots, p) = P(\text{Datos}, p)}}{\underbrace{\sum_{p, a_0 \ldots a_T} P(r_0, c_0, s_0, a_0, \ldots, p)}_{P(r_0, c_0, s_0, \ldots) = P(\text{Datos})}} \tag{4}$$

En particular, si revisamos el numerador veremos que lo podemos reescribir de la siguiente forma.

$$\begin{aligned} \sum_{a_0 \ldots a_T} P(r_0, c_0, s_0, a_0, \ldots, p) &\overset{3}{=} \sum_{a_0 \ldots a_T} \left(P(p) \prod_{t=0}^{T} P(r_t, c_t, s_t, a_t|p) \right) \\ &= P(p) \prod_{t=0}^{T} \underbrace{\sum_{a_t} P(r_t, c_t, s_t, a_t|p)}_{P(r_t, c_t, s_t|p)} = \underbrace{P(p)}_{\text{Prior}} \underbrace{\prod_{t=0}^{T} P(r_t, c_t, s_t|p)}_{\text{Verosimilitud}} \end{aligned} \tag{5}$$

Es decir, cuando calculamos el posterior de la memoria p vemos que en la verosimilitud los eventos son independientes entre sí. Cuando estudiemos el flujo de inferencia en estructuras causal vamos

a ver un método más simple para sacar este tipo de conclusiones, en el que se descompone las reglas de la probabilidad como mensajes que se envían los nodos de la red causal (factor graph).

$$P(p|\text{Datos} = \{(c_0, s_0, r_0), (c_1, s_1, r_1), \dots\}) = \frac{\overbrace{\prod_t P(c_t, s_t, r_t|p)}^{\text{Verosimilitud}} \overbrace{P(p)}^{\text{Prior}}}{\sum_p P(p) \prod_t P(c_t, s_t, r_t|p)} \tag{6}$$

Este posterior va a ser importante a la ahora de predecir lo que ocurre con el siguiente evento, lo que haremos en la siguiente sección. Si bien la probabilidad de acordarse puede es una variable continua y es posible encontrar una solución proporcional simple (pues hay solo 3 valores posibles para el likelihood), a efectos prácticos es suficiente que evalúen un conjunto finito de valores, de al menos 21 valores desde 0 a 1 equidistantes.

2.2. Calcular la predicción de un episodio dado los datos de los episodios anteriores

Para calcular la predicción del siguiente episodio dada la información de los eventos anteriores vamos a usar el último posterior de p como priori para el nuevo episodio.

$$P(\text{Episodio}_T = (c_T, s_T, r_T)|\text{Datos}_{0:T-1} = \{(c_0, s_0, r_0), \dots, (c_{T-1}, s_{T-1}, r_{T-1})\})$$

Recordar que cualquier conclusión que necesitemos alcanzar la podemos obtener aplicando las dos reglas de la probabilidad: la regla de la suma (marginal) y la regla del producto (condicional). En particular, en probabilidad las predicciones se realizan con la contribución de todas las hipótesis.

$$P(c_T, s_T, r_T|\text{Datos}_{0:T-1}) = \sum_p \sum_{a_T} P(c_T, s_T, r_T, a_T, p|\text{Datos}_{0:T-1})$$

Siguiendo la misma linea de razonamiento que hicimos en la sección anterior, podemos llegar a la conclusión que,

$$P(c_T, s_T, r_T|\text{Datos}_{0:T-1}) = \sum_p \sum_{a_T} P(r_T)P(c_T)P(s_T|r_T, c_T, a_T)P(a_T|p)P(p|\text{Datos}_{0:T-1}) \tag{7}$$

Cuando estudiemos el flujo de inferencia en estructuras causales vamos a ver lo simple que se vuelve alcanzar este tipo de conclusiones aplicando el algoritmo suma-producto, que descompone la aplicación de las reglas de la probabilidad como mensajes que se envían los nodos de las red causal (factor graph). Acá el punto importante es no olvidarse de utilizar el último posterior, $P(p|\text{Datos}_{0:T-1})$ para predecir el siguiente evento T.

2.3. Calcular la predicción que hace el modelo alternativo M_A sobre todo el conjunto de datos.

Calcular la verosimilitud del modelo alternativo como el producto de las predicciones de cada uno de los episodios dado los episodios anteriores.

$$P(\text{Datos}_{0:T}|M_A) = P(\text{Episodio}_0|M_A)P(\text{Episodio}_1|\text{Datos}_0, M_A)P(\text{Episodio}_2|\text{Datos}_{0:1}, M_A)\dots$$

La predicción sobre un conjunto de datos grandes necesariamente resulta ser un número muy cercano a 0. Esto ocurre porque los elementos del producto son probabilidades, números entre 0 y 1, por lo que a medida que vamos agregando episodios este número se va acercando tanto al cero que deja de poder ser representado por una computadora. Para poder expresarlo en una computadora, vamos a calcular el exponente asociado a ese número, que crece a una velocidad exponencialmente más lenta.

$$\log_{10} P(\text{Datos}_{0:T}|M_A) = \log_{10} P(\text{Episodio}_0|M_A) + \log_{10} P(\text{Episodio}_1|\text{Datos}_0, M_A) + \dots$$

2.4. Comparar el desempeño del modelo alternativo respecto de los modelos Base y el modelo MontyHall.

Cuando trabajamos con el exponente de las predicciones no vamos a poder calcular el posterior de los modelos directamente. En estos casos, para comparar el desempeño de los modelos lo que hacemos es comparar modelos de a pares.

$$\frac{P(M_i|\text{Datos})}{P(M_j|\text{Datos})} = \frac{P(\text{Datos}|M_i)P(M_i)}{P(\text{Datos}|M_j)P(M_j)} \stackrel{*}{=} \underbrace{\frac{P(\text{Datos}|M_i)}{P(\text{Datos}|M_j)}}_{\text{Bayes factor}} \tag{8}$$

Al dividir el valor a posteriori de los dos modelos alternativos i y j se cancela el denominador constante teorema de Bayes, $P(\text{Datos})$. Esa es la primera transformación que hacemos. Además, en el caso de que tengamos un prior uniforme entre modelos, $P(M_i) \stackrel{*}{=} P(M_j)$, también se cancelan los priors y la comparación del posterior de los modelos se reduce a la comparación de sus predicciones, $P(\text{Datos}|M)$. Este cociente entre predicciones de los modelos se conoce como *Bayes factor*.

Como hemos mencionado en la sección anterior, las predicciones sobre conjunto de datos relativamente conviene expresarla en escala logarítmica. Esta transformación tiene la ventaja adicional de hacer que el cociente sea simétrico. Por ejemplo, si comparamos en órdenes de magnitud el desempeño predictivo de los modelos Base M_0 y Monty Hall M_1 sobre los datos generados en el ejercicio anterior obtendremos una diferencia de aproximadamente 4 órdenes de magnitud.

$$\log_{10} \underbrace{\frac{P(\text{Datos}|M_1)}{P(\text{Datos}|M_0)}}_{\text{Bayes factor}} = \underbrace{\log_{10} P(\text{Datos}|M_1) - \log_{10} P(\text{Datos}|M_0)}_{\text{Diferencia predicitva en ordenes de magnitud}} \approx (-17) - (-21) = 4 \tag{9}$$

pues en el ejercicio anterior la predicción que el modelo base hizo del conjunto de datos era $P(\text{Datos}|M_0) \approx 3{,}37 \times 10^{-17}$ y la predicción que el modelo Monty Hall hizo era de $P(\text{Datos}|M_0) \approx 8{,}23 \times 10^{-21}$.

Para interpretar el significado del exponente del Bayes factor (9) es importante recordar que la verosimilitud de los modelos $P(\text{Datos}|M)$ funciona como filtro de la creencia previa. En base 10, una diferencia de un orden de magnitud significa que uno de los modelos preservó 10 veces más creencia que el otro, dos ordenes de magnitud significa que un modelos preservó 100 veces más creencia que el otro, y así sucesivamente. Aunque estos números parezcan extraordinarios, cuatro órdenes de magnitud se considera en el límite de una diferencia no concluyente. Cuando las bases de datos crecen, la diferencia en órdenes de magnitud continúan creciendo, por lo que es normal ver diferencia de 10000, pero en órdenes de magnitud! En esos casos, para ganar intuición es útil calcular la predicción "típica".

2.5. Calcular la predicción típica que hace el modelo de los episodios.

Dado que la predicción sobre un conjunto de datos se descompone como el producto de las predicciones, la predicción típica es su media geométrica, la raiz N-ésima de la predicción sobre todo el conjunto de datos (donde N es el tamaño del conjunto de datos $N = |\text{Datos}|$).

$$\text{Predicción típica} := \underbrace{P(\text{Datos} = \{d_1, d_2, \ldots, d_N\}|M))^{1/N}}_{\text{Media geométrica}} \tag{10}$$

Decimos que es típica justamente porque al reemplazar cada una de las predicciones individuales que componen en la secuencia de predicciones por el valor de la media geometrica volvemos a obtener exactamente el valor de la predicción sobre todo el conjunto de datos.

$$\begin{aligned} P(\text{Datos} = \{d_1, d_2, \ldots, d_N\}|M) &= P(d_1|M)P(d_2|d_1, M)\ldots \\ &= \prod_{i \in \{1,\ldots,N\}} \text{Predicción típica} = \text{Predicción típica}^N \end{aligned} \tag{11}$$

Para calcular la media geométrica vamos a enfrentar el mismo problema de representación computacional que señalamos respecto de las predicciones sobre conjuntos de datos grandes. Por eso, para calcularla hay que trabajar con el exponente de la predicción. La forma más sencilla es ir guardando la suma de los exponentes de las predicciones individuales, luego obtener el exponente promedio, y finalmente transformar ese exponente en número.

$$\begin{aligned} \text{Predicción típica} &= 10^{\overbrace{\log_{10}(P(d_1|M)P(d_2|d_1,M)\dots)^{1/N}}^{\text{Exponente de la predicción típica}}} \\ &= 10^{\frac{1}{N}\left(\log_{10} P(d_1|M)+\log_{10} P(d_2|d_1,M)+\dots\right)} \end{aligned} \tag{12}$$

Este número representa la predicción típica, una probabilidad que irá entre 0 y 1. Si calculan la predicción típica del modelo Base en los datos del ejercicio anterior verán que es de 0,382 y la del modelo Monty Hall de 0,454. Dado que observamos en total $N = 48$ datos (3 datos en cada una de los $T = 16$ episodios), podemos usar la predicción típica para recuperar la predicción conjunta.

$$P(\text{Datos}|M_0) = 0{,}382^{48} \qquad\qquad P(\text{Datos}|M_1) = 0{,}454^{48}$$

¿Por qué podemos decir que, en promedio (geométrico), el modelo base preserva solo el 38,2 % de la creencia previa luego de cada nueva observación, mientras que el modelo Monty Hall preserva 45,4 %?

2.6. Calcular el posterior en los primeros episodios y graficar

Para poder graficar cómo cambia la creencia de los modelos a medida que vamos observando episodios vamos a calcular el posterior de los tres modelos en los primeros 60 episodios. Debería quedar algo similar a lo siguiente.