


Examen integrador

Apuestas de vida

Inferencia Bayesiana Causal 1
2do cuatrimestre 2024
UBA - UNSAM

p= 

Laboratorios de
Métodos Bayesianos
Gustavo Landfried

Entrega: antes del miércoles 18 a las 23:59 horas.

1. Objetivo

El ejercicio de inferencia es una Monty Hall en el cual el sesgo de la persona que esconde el regalo cambia a lo largo del año y se repite todos los años. Para ello se provee un archivo `csv` con la cantidad de veces n_{ij} que se observó el regalo en una caja (columna i) en cierto día del año (fila j). El objetivo de esta tarea es crear un archivo de respuesta `csv`, con las mismas columnas y filas, pero que contenga la estimación de la probabilidad de la posición del regalo en la caja i el día j del año, $\mathbb{P}(r_j = i)$. La evaluación del examen de inferencia será el producto de todas las predicciones elevadas por la probabilidad correcta $P(r_j = i)$.

$$\mathcal{V} = \prod_j \prod_i \mathbb{P}(r_j = i)^{P(r_j = i)} \quad (1)$$

Además, el `csv` debe incluir una columna adicional que indique la caja que se decide elegir en cada día j del año, c_j . Tener en cuenta que la persona que da la pista no tiene sesgo, elige con misma probabilidad cualquiera de las cajas que están libres (cajas que no contienen el regalo y que no son la caja elegida en ese día c_j). La evaluación del problema de toma de decisiones será el producto de las distribuciones de probabilidad a posteriori sobre la posición del regalo para cada posible pista, s_j , elevada por el producto de la probabilidad a posteriori real de la posición del regalo multiplicada por la probabilidad de recibir la pista, s_j .

$$\mathcal{V} = \prod_j \prod_i \prod_{k \neq c_j} \mathbb{P}(r_j = i | s_j = k, c_j)^{P(r_j = i | s_j = k, c_j) P(s_j = k | c_j)} \quad (2)$$

2. Inferencia

En esta sección se dan algunas ideas para resolver el problema. En primer lugar empezamos pensando un problema más simple, visto en la unidad 1. Para inferir la probabilidad p de una

variable aleatoria binaria b podemos proponer un modelo conjugado Beta-Binomial.

$$P(p) = \text{Beta}(p|\alpha, \beta)$$

$$P(b|p) = \text{Bernoulli}(b|p)$$

Si en nuestra base de datos observamos N_0 realizaciones de $b = 0$ y N_1 realizaciones de $b = 1$, luego el posterior de p es

$$P(p|\text{Datos}) = \text{Beta}(\alpha + N_0, \beta + N_1)$$

Cuando la variable aleatoria no es binarias, como ocurre con la posición del regalo r en el problema Monty Hall, $r \in \{0, 1, 2\}$, se puede proponer un modelo conjugado Dirichlet-Catagórica.

$$P(\mathbf{p}) = \text{Dirichlet}(\mathbf{p}|\alpha_0, \alpha_1, \alpha_2)$$

$$P(r|\mathbf{p}) = \text{Categorical}(r|\mathbf{p})$$

donde \mathbf{p} es un vector de probabilidades. Si en nuestra base de datos observamos N_0 realizaciones de $r = 0$ y N_1 realizaciones de $r = 1$, y N_2 realizaciones de $r = 2$, luego el posterior de \mathbf{p} es $\text{Dirichlet}(\alpha_0 + N_0, \alpha_1 + N_1, \alpha_2 + N_2)$. En la siguiente figura se muestran el 2-simplex de tres distribución Dirichlet distintas.

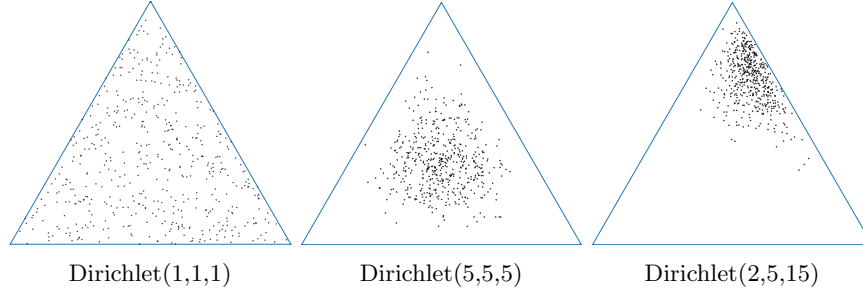


Figura 1. Distribuciones Dirichlet

Lo más sencillo para resolver este problema sería entonces proponer un modelo Dirichlet-Catagórico independiente para cada día del año y computar el posterior con la información específica a la cantidad de observaciones N_0 , N_1 y N_2 correspondientes a ese día, $\text{Dirichlet}(\alpha_0 + N_0, \alpha_1 + N_1, \alpha_2 + N_2)$.

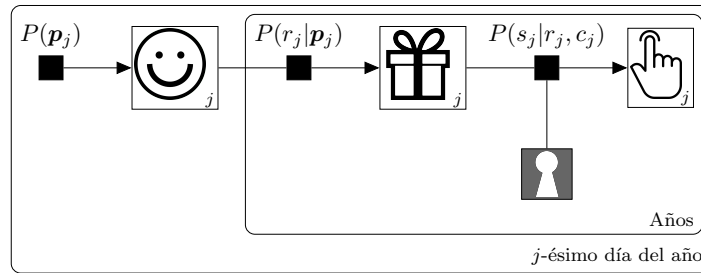


Figura 2. Modelo con sesgos independientes

Si bien el sesgo cambia, esto no significa que sea independiente del sesgo en los pasos temporales contiguos. Por el contrario, verán que el cambio en el sesgo es lento, es decir, el sesgo en un período temporal se parece mucho al sesgo en los períodos temporales contiguos. Para aprovechar esta información vamos a proponer un modelo de historia completa, en el que todos los días están

conectados entre sí por la influencia que el sesgo de un día tiene en el sesgo del día siguiente, donde el sesgo del día 365 afecta al sesgo del día 1, el sesgo del día 1 afecta al sesgo del día 2 y así sucesivamente.

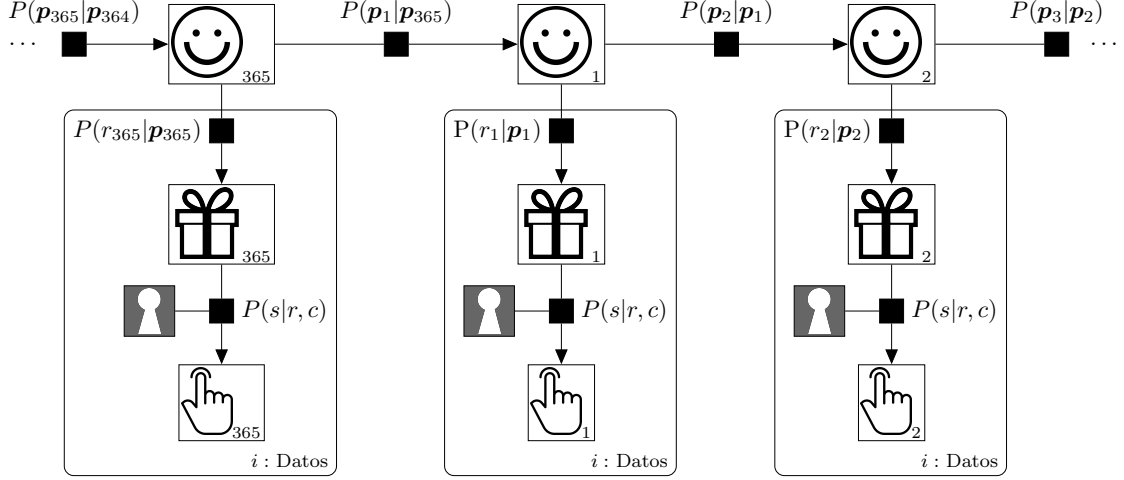


Figura 3. Especificación mediante factor graph del modelo Monty Hall con sesgo cíclico.

Notar que si el sesgo de la persona se repite todos los años, la especificación del modelo mediante factor graph tendrá naturalmente un ciclo. Esto contrasta con los modelos causales vistos en la materia, en los que se requería que fueran DAGs (Directed Acyclic Graphs). Una de las razones por las cuales se requiere que los modelos sean DAGs tiene que ver con preservar la relación causa-efecto. Pero además, solo los DAGs inducen descomposiciones válidas respecto de la regla de la cadena.

Sin embargo, es posible especificar modelos con ciclos y usar métodos eficientes de inferencia. En particular, el algoritmo *loopy belief propagation* visto en la unidad de series temporales de la materia, también alcanza convergencia respecto de las distribuciones marginales del modelo luego de una cantidad finita de iteraciones. En este ejercicio proponemos que realicen una simplificación y aproximen el posterior del sesgo mediante un promedio ponderado de las últimas estimaciones del sesgo de sus vecinos y la contribución de los datos.

$$p(\mathbf{p}) = \text{Dirichlet} \left(\lambda \left(\frac{\alpha_{0,t-1}}{2} + \frac{\alpha_{0,t+1}}{2} \right) + N_{0,t}, \dots \right) \quad (3)$$

Donde el parámetro λ controla la memoria respecto de los datos vecinos. Existen otras formas de resolver este problema. Elijan la que prefieran. Pueden usar Chat Bots para realizar el trabajo mientras entiendan lo que están haciendo. Lo único que importa es que las predicciones de los regalos reportadas en el csv sean lo más parecidas a las probabilidades reales con la que se generaron los datos. Notar que dado que el modelo es,

$$P(\mathbf{p}) = \text{Dirichlet}(\mathbf{p} | \alpha_0, \alpha_1, \alpha_2) \quad P(r | \mathbf{p}) = \text{Categorical}(r | \mathbf{p})$$

la predicción marginal de r es,

$$p(r = 0) = \frac{\alpha_0}{\alpha_0 + \alpha_1 + \alpha_2}$$

Suban en el csv estas probabilidades para cada caja (columna) y para cada día (fila). Por último, no se olviden de agregar una columna adicional en el csv con la caja que conviene elegir para recibir la pista.

No duden en comunicarse con su docente a cargo por el medio que sea. Envíen el csv, y el código con el que hicieron las estimaciones al correo electrónico gustavolandfried@gmail.com. Y si la consigna no es clara o tiene un error, no duden en comunicarse inmediatamente al mismo correo electrónico. Deben entregar esta tarea antes del miércoles 18 de diciembre a las 23:59 horas.