

7. 4. REDES BAYESIANAS.

1. Presentación intuitiva
2. Definición formal de red bayesiana
 - Algoritmo para comprobar independencias
3. Algoritmo de propagación de probabilidades en árboles
4. Evidencia virtual

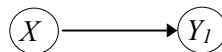
1. Presentación intuitiva¹.

Antes de presentar formalmente la teoría matemática de las redes bayesianas, explicaremos mediante ejemplos sencillos el significado intuitivo de los conceptos que después introduciremos.

En una red bayesiana, cada nodo corresponde a una variable, que a su vez representa una entidad del mundo real. Por tanto, de aquí en adelante hablaremos indistintamente de nodos y variables, y los denotaremos con letras mayúsculas, como X . Utilizaremos la misma letra en minúscula, x , para referirnos a un valor cualquiera de la variable X . Los arcos que unen los nodos indican relaciones de *influencia causal*.

Ejemplo 1.

La red bayesiana no trivial más simple que podemos imaginar consta de dos variables, que llamaremos X e Y_I , y un arco desde la primera hasta la segunda.



Para concretar el ejemplo, supongamos que X representa paludismo e Y_I representa gota gruesa, que es la prueba más habitual para detectar la presencia de dicha enfermedad.

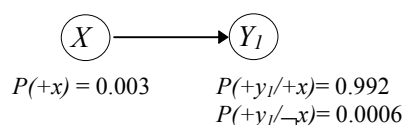
Cuando X sea una variable binaria, denotaremos por $+x$ la presencia de aquello a lo que representa y por $-x$ a su ausencia. Así, por ejemplo en este caso $+x$ significará “el paciente tiene paludismo” y $-x$ “el paciente no tiene paludismo”; $+y_I$ significará un resultado positivo del test de la gota gruesa y $\neg y_I$ un resultado negativo.

La información cuantitativa de una red bayesiana viene dada por:

- La probabilidad a priori de los nodos que no tienen padres.
- La probabilidad condicionada de los nodos con padres.

Por tanto, en nuestro ejemplo, los datos que debemos conocer son $P(x)$ y $P(y_I/x)$.

Así, la red bayesiana completa sería:



¹ Esta presentación intuitiva está tomada de los Apuntes de razonamiento aproximado de Francisco Javier Díez.

Veamos qué significado tienen en este caso estos valores:

- $P(+x) = 0.003$ indica que, a priori, un 0.3% de la población padece el paludismo. En medicina, esto se conoce como *prevalencia* de la enfermedad.
- $P(+y_I/+x) = 0.992$ indica que cuando hay paludismo, el test de la gota gruesa da positivo en el 99.2% de los casos. Esto se conoce como *sensibilidad* del test.
- $P(+y_I/-x) = 0.0006$ indica que cuando no hay paludismo, el test de la gota gruesa da positivo en el 0.06% de los casos, y negativo en el 99.94%. A esta segunda probabilidad se la llama *especificidad* del test.

En medicina siempre se buscan las pruebas con mayor grado de sensibilidad y especificidad.

Conociendo estos datos, podemos calcular:

a) La probabilidad a priori de Y_I ,

$$P(+y_I) = P(+y_I/+x) P(+x) + P(+y_I/-x) P(-x) = 0.00357.$$

$$P(-y_I) = P(-y_I/+x) P(+x) + P(-y_I/-x) P(-x) = 0.99643.$$

b) Las probabilidades a posteriori dada una evidencia observada e , $P^*(x) = P(x/e)$.

Supongamos que el test de la gota gruesa ha dado positivo. ¿Qué probabilidad hay ahora de que la persona padezca la enfermedad?. Si la prueba tuviese fiabilidad absoluta, esta probabilidad sería del 100%. Pero como existe la posibilidad de que haya habido un falso positivo, buscamos $P^*(+x) = P(+x/+y_I)$. Para calcularla, podemos aplicar el teorema de Bayes:

$$P^*(+x) = P(+x/+y_I) = \frac{P(+x) P(+y_I / +x)}{P(+y_I)} = \frac{0.003 \cdot 0.992}{0.00357} = 0.83263$$

Es decir, de acuerdo con el resultado de la prueba, hay un 83,2% de probabilidad de que el paciente tenga paludismo.

De la misma forma podríamos calcular $P^*(-x)$:

$$P^*(-x) = P(-x/+y_I) = \frac{P(-x) P(+y_I / -x)}{P(+y_I)} = \frac{0.997 \cdot 0.0006}{0.00357} = 0.16737$$

Que, por supuesto, es la probabilidad complementaria.

La expresión general del teorema de Bayes que hemos utilizado es:

$$P^*(x) = P(x/y) = \frac{P(x) P(y / x)}{P(y)}$$

Por razones que quedarán claras más adelante, vamos a reescribirla como;

$$P^*(x) = \alpha P(x) \lambda_{Y_I}(x)$$

Donde $\alpha = [P(y)]^{-1}$ y $\lambda_{Y_I}(x) = P(y/x)$.

Con la fórmula expresada de esta forma, queda claro que la probabilidad a posteriori de la variable X depende fundamentalmente de la probabilidad a priori de X (prevalencia de la enfermedad) y de la probabilidad condicionada de Y dado X (sensibilidad y especificidad del test), puesto que α juega simplemente el papel de una constante de normalización.

Utilizando esta nueva expresión, podemos repetir los cálculos:

$$P^*(+x) = \alpha 0.003 0.992 = 0.00298 \alpha.$$

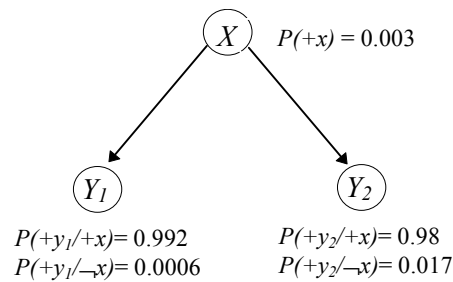
$$P^*(-x) = \alpha 0.997 0.0006 = 0.000598 \alpha.$$

Y normalizando obtenemos el mismo resultado que antes.

Para el caso en que el test de la gota gruesa diese negativo, la probabilidad a posteriori de padecer paludismo se calcula con un procedimiento totalmente análogo.

Ejemplo 2.

Supongamos que ampliamos el modelo anterior añadiendo un nuevo efecto del paludismo, la fiebre, que representaremos mediante la variable Y_2 . La red bayesiana se modifica entonces y queda como se muestra en la figura:



Vemos aquí que, para el paludismo, la fiebre tiene menor especificidad que la gota gruesa. Así, este sencillo modelo tiene en cuenta que hay muchas otras causas que pueden producir fiebre.

Veamos qué tipo de conclusiones podemos extraer a partir de esta información.

a) Supongamos que $e = \{+y_2\}$. Entonces, podemos calcular como antes la probabilidad a posteriori de que el paciente tenga paludismo sabiendo que tiene fiebre:

$$P^*(+x) = P(+x/+y_2) = \alpha 0.003 0.98 = 0.00294 \alpha = 0.148$$

$$P^*(-x) = P(-x/+y_2) = \alpha 0.997 0.0017 = 0.016949 \alpha = 0.852.$$

Como podemos observar hay solo un 0,148 de probabilidad de que tenga paludismo, resultado mucho más bajo que antes.

b) Supongamos que $e = \{+y_1, +y_2\}$. ¿Cuál es ahora $P^*(x) = P(x/+y_1, +y_2)$?

Para calcularla, usamos de nuevo el teorema de Bayes:

$$P^*(x) = P(x/y_1, y_2) = \frac{P(x) P(y_1, y_2 / x)}{P(y_1, y_2)}$$

Pero ahora vemos que hay datos del problema que no conocemos, como $P(y_1, y_2)$ y $P(y_1, y_2/x)$. Para poder seguir nuestros cálculos, necesitamos realizar unas hipótesis adicionales que se llaman *hipótesis de independencia condicional*. En concreto, vamos a suponer que las variables Y_1 e Y_2 son independientes dados su padre común, X , es decir:

$$P(y_1, y_2/x) = P(y_1/x) P(y_2/x).$$

Si suponemos esto podremos continuar con los cálculos porque $P(y_1, y_2)$ se obtendrá como constante de normalización.

¿Qué significa aceptar esta hipótesis?. Significa aceptar que, conocido que un paciente tiene paludismo, el hecho de que tenga fiebre o no, no depende de que el resultado del test de la gota gruesa sea positivo o negativo, lo cual parece razonable.

Para seguir con la nueva formulación que introdujimos en el ejemplo 1, vamos a denotar por $\lambda(x)$ a $\lambda_{Y_1}(x) \lambda_{Y_2}(x)$. Entonces tendríamos que

$$P^*(x) = \alpha P(x) \lambda(x).$$

En nuestro ejemplo, $e = \{+y_1, +y_2\}$, y por tanto;

$$\lambda(+x) = \lambda_{Y_1}(+x) \lambda_{Y_2}(+x) = 0.97216$$

$$\lambda(-x) = \lambda_{Y_1}(-x) \lambda_{Y_2}(-x) = 0.0337.$$

Por tanto,

$$P^*(+x) = 0.9663$$

$$P^*(-x) = 0.0337$$

Como era de esperar, cuando tenemos dos evidencias en favor del paludismo, la probabilidad resultante es mayor que la correspondiente a cada uno de ellos por separado.

c) En el caso en que tengamos un hallazgo a favor y otro en contra, podemos ponderar su influencia mediante estas mismas expresiones. Supongamos por ejemplo que el test de la gota gruesa ha dado negativo, pero que el paciente padece fiebre, esto es, $e = \{-y_1, +y_2\}$, y calculemos la probabilidad de que el paciente padezca fiebre.

$$\lambda(+x) = \lambda_{Y_1}(+x) \lambda_{Y_2}(+x) = P(-y_1/+x) P(+y_2/+x) = 0.008 \cdot 0.98 = 0.00784$$

$$\lambda(-x) = \lambda_{Y_1}(-x) \lambda_{Y_2}(-x) = P(-y_1/-x) P(+y_2/-x) = 0.994 \cdot 0.017 = 0.01699$$

Vemos que hay más evidencia a favor de $-x$ que a favor de $+x$, debido principalmente a la alta sensibilidad de la prueba de la gota gruesa. Al tener en cuenta además la probabilidad a priori de la enfermedad, nos queda que,

$$P^*(+x) = 0.0014$$

$$P^*(-x) = 0.9986$$

d) Aún podemos extraer más información de este ejemplo. Supongamos ahora que la que tenemos un paciente con fiebre al que aún no hemos realizado el test de la gota gruesa, es decir, la evidencia considerada es $e = \{+y_2\}$. ¿Cuál es la probabilidad de que al hacerle el test, el resultado sea positivo?.

Buscamos ahora $P(y_1/y_2)$. Por teoría elemental de probabilidad, sabemos que;

$$P^*(y_1) = P(y_1/y_2) = \sum_x P(y_1/x, y_2) P(x/y_2) = \sum_x P(y_1/x, y_2) \frac{P(x, y_2)}{P(y_2)}.$$

Aplicando la hipótesis de independencia condicional y llamando

$$\pi_{Y_1}(x) = P(x, y_2) = P(x) P(y_2/x)$$

$$\alpha = [P(y_2)]^I,$$

la expresión anterior nos queda:

$$P^*(y_1) = \alpha \sum_x P(y_1/x) \pi_{Y_1}(x).$$

Sustituyendo los valores numéricos de nuestro ejemplo, tenemos que,

$$\pi_{Y_1}(+x) = P(+x)P(+y_2/+x) = 0.003 \cdot 0.98 = 0.00294$$

$$\pi_{Y_1}(-x) = P(-x)P(+y_2/-x) = 0.997 \cdot 0.017 = 0.01695$$

Y, finalmente,

$$P^*(+y_1) = \alpha [\pi_{Y_1}(+x) P(+y_1/+x) + \pi_{Y_1}(-x) P(+y_1/-x)] = 0.14715$$

$$P^*(-y_1) = \alpha [\pi_{Y_1}(+x) P(-y_1/+x) + \pi_{Y_1}(-x) P(-y_1/-x)] = 0.85285$$

Resulta interesante comparar las expresiones utilizadas para calcular la probabilidad a priori $P(y_1)$ y la a posteriori $P^*(y_1)$. Para la primera, utilizábamos $P(x)$, ahora hemos utilizado $\pi_{Y_1}(+x)$, que indica la probabilidad de x tras considerar la evidencia relativa a x diferente de Y_1 .

Vemos así como la información que aporta el nodo Y_2 modifica la probabilidad de X , y, en consecuencia, también la de Y_1 . El carácter simultáneamente ascendente y descendente del mecanismo de propagación es lo que nos permite utilizar la red tanto para realizar inferencias abductivas (cuál es el diagnóstico que mejor explica los hallazgos) como predictivas (cuál es la probabilidad de obtener cierto resultado en el futuro). Un mismo nodo puede ser tanto fuente de información como objeto de predicción, dependiendo de cuáles sean los hallazgos disponibles y el objeto del diagnóstico.

Terminada ya esta presentación intuitiva, vamos a introducir formalmente las redes bayesianas.

2. Definición formal de red bayesiana.

Antes de definir formalmente las redes bayesianas, vamos a definir algunos conceptos de teoría de grafos y teoría de la probabilidad:

Definiciones previas

- **Arco.** Es un par ordenado (X, Y) . Esta definición de arco corresponde a lo que en otros lugares se denomina *arco dirigido*. En la representación gráfica, un arco (X, Y) viene dado por una flecha desde X hasta Y .

- **Grafo dirigido.** Es un par $G = (N, A)$ donde N es un conjunto de nodos y A un conjunto de arcos definidos sobre los nodos.
 - **Grafo no dirigido.** Es un par $G = (N, A)$ donde N es un conjunto de nodos y A un conjunto de arcos no orientados (es decir, pares no ordenados (X, Y)) definidos sobre los nodos.
 - **Camino.** Es una secuencia ordenada de nodos $(X_{i_1}, \dots, X_{i_r})$ tal que $\forall j = 1, \dots, r-1$, ó bien el arco $X_j \rightarrow X_{j+1} \in A$ o bien el arco $X_{j+1} \rightarrow X_j \in A$.
 - **Camino dirigido.** Es una secuencia ordenada de nodos $(X_{i_1}, \dots, X_{i_r})$ tal que para todo $j = 1, \dots, r-1$ el arco $X_j \rightarrow X_{j+1} \in A$.
 - **Ciclo:** es un camino no dirigido que empieza y termina en el mismo nodo X .
 - **Grafo acíclico:** es un grafo que no contiene ciclos.
 - **Padre.** X es un *padre* de Y si y sólo si existe un arco $X \rightarrow Y$. Se dice también que Y es **hijo** de X . Al conjunto de los padres de X se representa como $pa(X)$, y al de los hijos de X por $S(X)$.
 - **Antepasado o ascendiente.** X es un *antepasado* o ascendiente de Z si y sólo si existe un camino dirigido de X a Z .
 - **Conjunto ancestral** de un nodo X es un conjunto que contiene a X y a todos sus antepasados.
 - **Descendiente.** Z es un *descendiente* de X si y sólo si X es un antepasado de Z . Al conjunto de los descendientes de X lo denotaremos por $de(X)$.
 - **Variable proposicional** es una variable aleatoria que toma un conjunto exhaustivo y excluyente de valores. La denotaremos con letras mayúsculas, por ejemplo X , y a un valor cualquiera de la variable con la misma letra en minúscula, x .
- La definición de variable proposicional tiene su importancia a la hora de modelar un problema con una red bayesiana, ya que deberemos tener en cuenta que los nodos de la red son variables proposicionales y por tanto deben tomar un conjunto exhaustivo y excluyente de valores. Así si por ejemplo estamos construyendo un sistema de diagnóstico médico en que las enfermedades posibles son gripe, faringitis y alergia, cada una de estas enfermedades será representada por un nodo diferente (que tomará valores *si/no*) ya que nada impide que un paciente padezca dos o más enfermedades a la vez (es decir, el conjunto de valores no es *excluyente*), o una enfermedad no considerada por el sistema (no es *exhaustivo*). Sin embargo si estamos construyendo un sistema de clasificación de animales en el que hemos representado todas las posibilidades, debemos introducir una única variable, ya que un animal no puede ser a la vez una vaca y un burro.
- Dos variables X e Y son **independientes** si se tiene que $P(X/Y) = P(X)$. De esta definición se tiene una caracterización de la independencia que se puede utilizar como definición alternativa: X e Y son independientes sí y sólo sí $P(X, Y) = P(X) \cdot P(Y)$.

- Dos variables X e Y son **independientes** dado una tercera variable Z si se tiene que $P(X/Y,Z) = P(X/Y)$. De esta definición se tiene una caracterización de la independencia que se puede utilizar como definición alternativa: X e Y son independientes dado Z si y sólo si $P(X,Y/Z) = P(X/Z) \cdot P(Y/Z)$. También se dice que Z *separa condicionalmente* a X e Y .

Definición

Una red bayesiana es:

- Un conjunto de variables proposicionales, V .
- Un conjunto de relaciones binarias definida sobre las variables de V , E .
- Una distribución de probabilidad conjunta sobre las variables de V .

tales que:

- (V, E) forman un grafo acíclico, conexo y dirigido G .
- (G, P) cumplen las *hipótesis de independencia condicional*, también llamadas de *separación direccional*, que se enuncian a continuación.

Hipótesis de independencia condicional o separación direccional

Un grafo acíclico conexo y dirigido $G = (V, E)$ y una distribución de probabilidad conjunta P definida sobre las variables del grafo se dice que cumplen las hipótesis de independencia condicional o separación direccional, si

$\forall X \in V$ y $\forall Y \in V - \{X \cup de(X) \cup pa(X)\}$ se tiene que X es independiente de Y dado $pa(X)$
(es decir, $P(X/pa(X), Y) = P(X/pa(X))$)

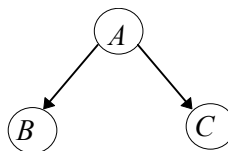
donde $de(X)$ denota al conjunto de descendientes, directos e indirectos, de X .

Así, para que se cumplan las condiciones de independencia condicional se debe cumplir que para toda variable X de V el conjunto de los padres directos de X debe *separar condicionalmente* a X de todo otro nodo Y de la red que no sea ni X , ni ninguno de sus descendientes, ni ninguno de sus padres (o lo que es lo mismo, X e Y deben ser independientes dado $pa(X)$).

Vamos a hacer un ejemplo en el que demostraremos que una red es una red bayesiana.

Ejemplo 3

Consideremos la red dada en la siguiente figura



En el que las variables que aparecen son binarias, junto con la siguiente distribución de probabilidad conjunta:

$$P(a_1, b_1, c_1) = 0.084$$

$$P(a_1, b_2, c_1) = 0.126$$

$$P(a_2, b_1, c_1) = 0.084$$

$$P(a_1, b_1, c_2) = 0.036$$

$$P(a_1, b_2, c_2) = 0.054$$

$$P(a_2, b_2, c_1) = 0.056$$

$$P(a_2, b_1, c_2) = 0.336$$

$$P(a_2, b_2, c_2) = 0.224$$

¿Es esto una red bayesiana?

Vemos que el grafo es acíclico, conexo y dirigido, así que tendríamos que probar que se satisfacen las hipótesis de independencia condicional.

- Para $X = A$.

$$pa(A) = \emptyset$$

$$V - \{A, \cup de(A) \cup pa(A)\} = \emptyset$$

luego no sería necesario hacer ninguna comprobación.

- Para $X = B$.

$$pa(B) = A.$$

$$V - \{B, de(B) \cup pa(B)\} = C.$$

Es decir, que B y C son independientes dado su padre común, A . Nótese que esta red presenta la misma estructura que la del ejemplo 2, y la hipótesis que ahora tenemos que demostrar es la misma que tuvimos que hacer allí para poder seguir con los cálculos.

Para probar esto, tendríamos que ver que

$$P(b_1/c_1, a_1) = P(b_1/c_2, a_1) = P(b_1/a_1)$$

$$P(b_1/c_1, a_2) = P(b_1/c_2, a_2) = P(b_1/a_2)$$

Puesto que las restantes comprobaciones no sería necesario hacerlas ya que al ser B una variable binaria, las probabilidades relativas a b_2 son complementarias de estas.

A modo de ejemplo vamos a comprobar una de ellas.

$$P(b_1/c_1, a_1) = \frac{P(a_1, b_1, c_1)}{P(a_1, c_1)} = \frac{P(a_1, b_1, c_1)}{P(a_1, b_1, c_1) + P(a_1, b_2, c_1)} = \frac{0.084}{0.084 + 0.126} = 0.4$$

$$\begin{aligned} P(b_1/a_1) &= \frac{P(a_1, b_1)}{P(a_1)} = \frac{P(a_1, b_1, c_1) + P(a_1, b_1, c_2)}{P(a_1, b_1, c_1) + P(a_1, b_2, c_1) + P(a_1, b_1, c_2) + P(a_1, b_2, c_2)} = \\ &= \frac{0.084 + 0.036}{0.084 + 0.126 + 0.036 + 0.054} = 0.4. \end{aligned}$$

Realizando las comprobaciones restantes veríamos que en este caso sí tenemos una red bayesiana.

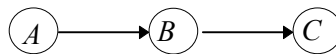
- Por último, para $X = C$, al ser la red simétrica, la única comprobación que tendríamos que hacer es ver que C es independiente de B dado su padre común, A , pero al ser la relación de independencia condicional simétrica, tendríamos que se satisfacen también las relaciones de independencia condicional.

Para este tipo de estructura se dice que los enlaces convergen *cola-con-cola* en el nodo A . De este modo, conocer el valor del padre común (A) *cierra la comunicación* entre los hijos (B y C), es decir, conocer el valor que toma un hijo ya no aporta información sobre el valor que puede tomar el otro.

Este tipo de estructura es el que teníamos en el ejemplo 2, en el que el paludismo tenía influencia causal en el test de la gota-gruesa y en la fiebre. Así, antes de saber con seguridad si una persona padece paludismo, conocer el resultado del test de la gota gruesa cambia mi opinión acerca de si padece paludismo, y a su vez esto cambia la probabilidad de que tenga fiebre. Sin embargo, una vez que sabemos que una persona padece paludismo, saber el resultado del test de la gota-gruesa ya no me aporta información sobre si tendrá o no fiebre.

Otros ejemplos,

- Para la red ;



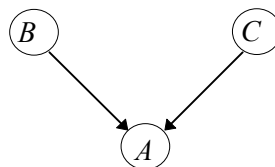
Habría que comprobar que A y C son independientes dado B

En este tipo de estructura se dice que los enlaces convergen *cola-con-cabeza* en B . Al conocer el valor de B , se *cierra la comunicación* entre el padre de B y el hijo de B .

Si por ejemplo tenemos que la metástasis (A) causa tumor cerebral (B) y jaquecas (C), a priori saber si tiene metástasis o no cambia mi opinión acerca de su probabilidad de desarrollar un tumor cerebral, lo que a su vez afecta a la probabilidad de padecer jaquecas. Sin embargo, una vez que se que la persona tiene un tumor cerebral, saber si ese tumor ha sido producido por una metástasis ya no afecta a la probabilidad de que la persona tenga jaquecas.

Otro ejemplo: supongamos que estamos haciendo un estudio sobre la bolsa. El estado de la bolsa ayer (A) influye en el estado de la bolsa hoy (B), que a su vez influye en el estado de la bolsa mañana (C). Por tanto, a priori el estado de la bolsa ayer tiene influencia causal en el estado de la bolsa mañana, pero esa influencia es sólo *a través* del estado de la bolsa hoy. En cuanto conocemos el valor de la bolsa hoy, la relación entre A y C se interrumpe, es decir, una variable ya no nos aporta información sobre otra. Los sistemas que presentan este comportamiento se dice que tienen la propiedad de Markov, que dice que, dado el presente, el futuro es independiente del pasado.

- Para la red ;



Habría que comprobar que B es independiente de C .

En este tipo de estructura se dice que los enlaces convergen *cabeza-con-cabeza* en A . Conocer el valor del hijo común (A) abre la comunicación entre los padres (B y C), ya que conocer el valor de un padre cambia las probabilidades del otro.

De este modo, pensemos en el caso en que hay dos enfermedades que provocan el mismo síntoma (por ejemplo, tanto la gripe como una infección de orina provocan fiebre).

A priori, las dos enfermedades son independientes (a menos que estén correladas, en cuyo caso aparecería un enlace de una a otra). Sin embargo, una vez que sabemos que el paciente padece una de las enfermedades queda explicado el síntoma, y por tanto la probabilidad de que padezca la otra enfermedad disminuye. Del mismo modo, si sabemos con certeza que no padece una de las enfermedades, la probabilidad de que padezca la otra aumenta. Este efecto se conoce con el nombre de *explaining-away*, que podríamos traducir como *descartar/potenciar causas*.

En la definición de red bayesiana, hemos partido de una distribución de probabilidad conjunta para las variables. Aparentemente, suponiendo que tuviésemos una red con N nodos y con variables binarias, haría falta conocer $2^N - 1$ valores. Sin embargo, las condiciones de independencia dadas por la separación direccional permiten que no sea necesario conocer todos estos valores, puesto que, como veremos en el siguiente Teorema, la distribución de probabilidad conjunta se puede expresar como producto de las distribuciones condicionadas de cada nodo dados sus padres.

Teorema (Factorización de la probabilidad).

Dada una red bayesiana, su distribución de probabilidad puede expresarse como:

$$P(x_1, \dots, x_n) = \prod_i P(x_i / pa(x_i)).$$

Demostración:

Es fácil construir una ordenación de las variables en la que los padres de cada nodo aparezcan siempre después de él. Supongamos por tanto que la ordenación $\{X_1, \dots, X_n\}$ cumple dicha propiedad. Por tanto:

$$P(x_1, \dots, x_n) = \prod_i P(x_i / x_{i+1}, \dots, x_n).$$

Pero por la forma de escoger la ordenación, el conjunto $\{X_{i+1}, \dots, X_n\}$ incluye a todos los padres de X_i , y, en consecuencia, la separación direccional nos dice que

$$P(x_i / x_{i+1}, \dots, x_n) = P(x_i / pa(x_i))$$

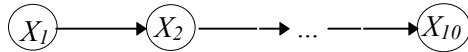
Con lo que concluimos la demostración.

La importancia de este teorema es que nos permite describir una red bayesiana a partir de la probabilidad condicionada de cada nodo (o la probabilidad a priori en el caso de nodos sin padres) en lugar de dar la probabilidad conjunta, que,

- requiere un número de parámetros exponencial en el número de nodos.
- plantea el problema de verificar la separación direccional.

Sin embargo, el número de parámetros requerido para dar las probabilidades condicionadas es mucho menor (proporcional al número de nodos), nos permite reconstruir la distribución conjunta aplicando el teorema, y además, a la hora de pedirle estos valores al experto, son valores con pleno significado, como vimos en el ejemplo 1.

Por ejemplo, para la red bayesiana dada por:



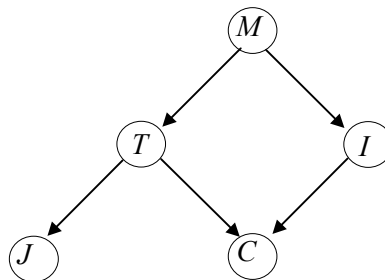
Suponiendo que todas las variables fuesen binarias, para dar la distribución conjunta habría que dar $2^{10}-1$ valores, sin embargo, si construimos la distribución conjunta a partir de los 19 valores necesarios para dar las condicionadas, tendremos además asegurado que se satisfacen las hipótesis de independencia condicional.

Ejemplo 4:

En un sistema de diagnóstico médico, supongamos que tenemos la siguiente información:

- Metástasis (M) causa tumor cerebral (T) e incremento en los niveles de calcio (I).
- Tumor cerebral causa coma (C).
- Incremento en nivel de calcio causa coma.
- Tumor cerebral causa fuertes jaquecas (J)

Representamos dicha información en la siguiente red bayesiana:



¿Qué independencias implica la red?

- Para M, tenemos una igualdad trivial $P(M)=P(M)$.
- Para I, $P(I/M) = P(I/M,T,C)$, es decir, I es independiente de todos los subconjuntos de $\{M,T,J\}$ dado M. Así por ejemplo dado metástasis, el incremento de calcio no depende de si hay o no tumor cerebral.
- Para T, $P(T/M) = P(T/M,I)$, es decir, T independiente de I dado M.
- Para C, $P(C/T,I) = P(C/T,I,M,J)$, es decir, C es independiente de todos los subconjuntos de $\{M,J\}$ dados T y I.
- Para J, $P(J/T) = P(J/T,M,I,C)$, es decir, J es independiente de todos los subconjuntos de $\{M,I,C\}$ dado T.

Según el teorema de factorización, los únicos datos que debemos pedir al experto son:

$P(m_1) = 0.2$	
$P(i_1/m_1) = 0.8$	$P(i_1/m_2) = 0.2$
$P(t_1/m_1) = 0.2$	$P(t_1/m_2) = 0.05$
$P(c_1/i_1,t_1) = 0.8$	$P(c_1/i_1,t_2) = 0.9$
$P(c_1/i_2,t_1) = 0.7$	$P(c_1/i_2,t_2) = 0.05$
$P(j_1/t_1) = 0.8$	$P(j_1/t_2) = 0.6$

Y a partir de estas probabilidades podríamos construir la distribución conjunta. Por ejemplo:

$$P(m_1, i_2, t_1, j_2, e_2) = P(m_1) \cdot P(i_2/m_2) \cdot P(t_1/m_1) \cdot P(c_2/i_2, t_1) \cdot P(j_2/t_1) = 0.2 \cdot 0.8 \cdot 0.2 \cdot 0.3 \cdot 0.2 = 0.00192.$$

Con la distribución conjunta P así construida tenemos una red causal que representa a la perfección la noción humana de causalidad. Una vez obtenida la distribución conjunta, podemos a partir de ella obtener la distribución de probabilidad que queramos. Por ejemplo, si queremos saber $P(M/C, J)$:

$$P(M/C=c_1, J=j_2) = \frac{P(M, c_1, j_2)}{P(c_1, j_2)} = \frac{\sum_{I, T} P(M, I, T, j_2, c_1)}{\sum_{M, I, T} P(M, I, T, j_2, c_1)}$$

Como vemos, una vez conocida la distribución conjunta es posible conocer la distribución de cualquier conjunto de variables, condicionadas o no condicionadas. Pero este método es computacionalmente demasiado costoso. Para evitar tener que realizar los cálculos de esta forma, se han desarrollado varios algoritmos más eficientes de propagación de probabilidades, que vamos a ver a continuación.

3. Algoritmo de propagación de probabilidades².

El teorema de factorización proporciona un primer método de actualización de las probabilidades dada la evidencia disponible, puesto que a partir de las probabilidades condicionadas es posible obtener la probabilidad conjunta, y a partir de ésta aplicando la definición de probabilidad condicionada y marginalizando la distribución conjunta sobre el conjunto de las variables de interés es posible conseguir $P(X/Y)$ donde X es cualquier subconjunto de V e Y cualquier conjunto de evidencias. Sin embargo, este procedimiento es computacionalmente costoso. Se han desarrollado muchos otros algoritmos más eficientes para el cálculo de las probabilidades, de los cuales sólo vamos a explicar en profundidad el algoritmo para redes con forma de árbol. El algoritmo consta de dos fases:

Fase de inicialización.

En esta fase se obtienen las probabilidades a priori de todos los nodos de la red, obteniendo un estado inicial de la red que denotaremos por S_0 .

Fase de actualización.

Cuando una variable se instancia, se actualiza el estado de la red, obteniéndose las probabilidades a posteriori de las variables de la red basadas en la evidencia considerada, adoptando la red un estado que denotaremos por S_I .

Este paso se repite cada vez que una variable se instancia, obteniéndose los sucesivos estados de la red.

La idea principal en la que se basa el algoritmo es la siguiente:

² Tomado de [Neapolitan, 1990]

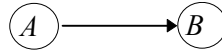
Cada vez que una variable se instancia, informa a sus nodos vecinos mediante el paso de lo que llamaremos mensajes, de la siguiente forma:

- La variable envía a su padre un mensaje, que llamaremos el λ -mensaje, para informarle de que ha cambiado su valor.
- La variable envía a todos sus hijos un mensaje, que llamaremos el π -mensaje, para informarlos de que ha cambiado su valor.

Así, la información se va propagando por la red tanto en sentido ascendente como descendente.

Estos mensajes asignan a cada variable unos valores que llamaremos λ -valor y π -valor. Multiplicando estos valores obtendremos las probabilidades a posteriori de cada una de las variables de la red.

Tanto los valores como los mensajes son vectores de números. Por ejemplo, supongamos que tenemos el arco:



en el que la variable A toma tres valores posibles a_1, a_2, a_3 , y la variable B toma dos, b_1 y b_2 , tendríamos que:

- Si B se instancia, enviará un λ -mensaje a A , $\lambda_B(A) = (\lambda_B(a_1), \lambda_B(a_2), \lambda_B(a_3))$.
- Si A se instancia, enviará un π -mensaje a B , $\pi_B(A) = (\pi_B(a_1), \pi_B(a_2), \pi_B(a_3))$.

En función de esos mensajes, tendremos un λ -valor y π -valor para A ,

$$\lambda(A) = (\lambda(a_1), \lambda(a_2), \lambda(a_3)) \quad \text{y} \quad \pi(A) = (\pi(a_1), \pi(a_2), \pi(a_3))$$

Y también un λ -valor y π -valor para B ,

$$\lambda(B) = (\lambda(b_1), \lambda(b_2)) \quad \text{y} \quad \pi(B) = (\pi(b_1), \pi(b_2))$$

Multiplicando los valores y normalizando, obtendremos las probabilidades asociadas a A o a B , según sea el caso.

Pasamos entonces a describir el algoritmo;

Fórmulas para el cálculo de λ y π -mensajes, λ y π -valores y probabilidades P^* :

1. Si B es un hijo de A , B tiene k valores posibles y A m valores posibles, entonces para $j=1,2,\dots,m$, el λ -mensaje de B a A viene dado por;

$$\lambda_B(a_j) = \sum_{i=1}^k P(b_i / a_j) \lambda(b_i).$$

2. Si B es hijo de A y A tiene m valores posibles, entonces para $j=1,2,\dots,m$, el π -mensaje de A a B viene dado por;

$$\pi_B(a_j) = \begin{cases} \pi(a_j) \prod_{\substack{c \in s(A) \\ c \neq B}} \lambda_c(a_j) & \text{si } A \text{ no ha sido instanciada (*)} \\ 1 & \text{si } A = a_j \\ 0 & \text{si } A \neq a_j. \end{cases}.$$

donde $s(A)$ denota al conjunto de hijos de A .

(*) Esta fórmula es válida en todos los casos. Hay otra fórmula cuya aplicación resulta a veces más sencilla, pero que sólo es válida cuando todas las probabilidades $P^*(a_i)$ son no nulas, que es:

$$P(a_j) / \lambda_B(a_j).$$

Esta fórmula da un π -mensaje distinto (pero proporcional al de la otra fórmula) e iguales probabilidades a posteriori.

3. Si B tiene k valores posibles y $s(B)$ es el conjunto de los hijos de B , entonces para $i=1,2,\dots,k$, el λ -valor de B viene dado por;

$$\lambda(b_i) = \begin{cases} \prod_{C \in s(B)} \lambda_C(b_i) & \text{si } B \text{ no ha sido instanciada} \\ 1 & \text{si } B = b_i \\ 0 & \text{si } B \neq b_i. \end{cases}$$

4. Si A es padre de B , B tiene k valores posibles y A tiene m valores posibles, entonces para $i=1,2,\dots,k$, el π -valor de B viene dado por;

$$\pi(b_i) = \sum_{j=1}^m P(b_i / a_j) \pi_B(a_j).$$

5. Si B es una variable con k posibles valores, entonces para $i = 1,2,\dots,k$ la probabilidad a posteriori basada en las variables instanciadas se calcula como:

$$P^*(b_i) = \alpha \lambda(b_i) \pi(b_i).$$

ALGORITMO:

1. Inicialización.

- A. Inicializar todos los λ -mensajes y λ -valores a 1.
- B. Si la raíz A tiene m posibles valores, entonces para $j = 1, \dots, m$, sea

$$\pi(a_j) = P(a_j).$$

- C. Para todos los hijos B de la raíz A , hacer

Enviar un nuevo π -mensaje a B usando la fórmula 2.

(En ese momento comenzará un flujo de propagación debido al procedimiento de actualización C).

Cuando una variable se instancia o una variable recibe un λ o π -mensaje, se usa uno de los siguientes procedimientos de actualización;

2. Actualización.

- A. Si una variable B se instancia a un valor b_j , entonces

BEGIN

- A.1. Inicializar $P^*(b_j) = 1$ y $P^*(b_i) = 0$, para todo $i \neq j$.
- A.2. Calcular $\lambda(B)$ usando la fórmula 3.
- A.3. Enviar un nuevo λ -mensaje al padre de B usando la fórmula 1.
- A.4. Enviar nuevos π -mensajes a los hijos de B usando la fórmula 2.

END

- B. Si una variable B recibe un nuevo λ -mensaje de uno de sus hijos y la variable B no ha sido instanciada todavía, entonces,

BEGIN

- B.1. Calcular el nuevo valor de $\lambda(B)$ usando la fórmula 3.
- B.2. Calcular el nuevo valor de $P^*(B)$ usando la fórmula 5.
- B.3. Enviar un nuevo λ -mensaje al padre de B usando la fórmula 1.
- B.4. Enviar nuevos π -mensajes a los otros hijos de B usando la fórmula 2.

END.

- C. Si una variable B recibe un nuevo π -mensaje de su padre y la variable B no ha sido instanciada todavía, entonces,

BEGIN

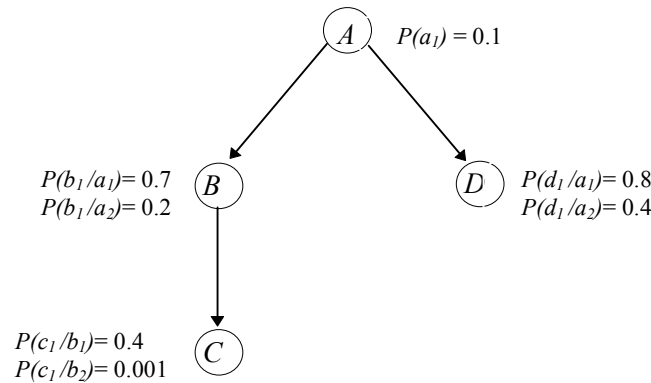
- C.1. Calcular el nuevo valor de $\pi(B)$ usando la fórmula 4.
- C.2. Calcular el nuevo valor de $P^*(B)$ usando la fórmula 5.
- C.3. Enviar nuevos π -mensajes a los hijos de B usando la fórmula 2.

END.

Para ilustrar el algoritmo de propagación de probabilidades, vamos a utilizar el siguiente ejemplo:

Ejemplo 4.

Supongamos que un señor piensa que su esposa le está siendo infiel. La red bayesiana que se construye para evaluar esta posibilidad es la siguiente:



donde la variable A se refiere a si la esposa está engañando al marido o no, la variable B se refiere a si la esposa cena con otro hombre o no, la variable C se refiere a si la esposa es vista cenando con otro hombre o no, y la variable D se refiere a si en el domicilio se reciben llamadas telefónicas sospechosas o no. Supondremos que la letra minúscula con el subíndice uno representa a la afirmación del hecho, y la minúscula con el subíndice 2, a la negación.

En primer lugar, vamos a calcular las probabilidades a priori de cada una de las variables de la red. Para hacer esto, también podríamos calcular la probabilidad conjunta como producto de las condicionadas, y luego sumar los términos necesarios para obtener cada probabilidad, pero este método no es computacionalmente factible en la mayoría de los casos. Vamos pues a hacerlo con el algoritmo.

Inicialización.

A. Ponemos todos los λ -mensajes y λ -valores a 1.

B. Hacemos $\pi(a_j) = P(a_j)$, para $j = 1, 2$. $\pi(A) = (0.1, 0.9)$.

C. A envía un mensaje a su hijo, B ,

$$\pi_B(a_1) = \pi(a_1) \lambda_D(a_1) = 0,1$$

$$\pi_B(a_2) = \pi(a_2) \lambda_D(a_2) = 0,9$$

B toma entonces nuevos π -valores;

$$\pi(b_1) = P(b_1/a_1) \pi_B(a_1) + P(b_1/a_2) \pi_B(a_2) = 0.7 \cdot 0.1 + 0.2 \cdot 0.9 = 0.25$$

$$\pi(b_2) = P(b_2/a_1) \pi_B(a_1) + P(b_2/a_2) \pi_B(a_2) = 0.75$$

Y con ellos y con los λ -valores de B , se obtienen las probabilidades:

$$P(b_1) = \alpha \cdot 0.25 \cdot 1 = 0.25.$$

$$P(b_2) = \alpha \cdot 0.75 \cdot 1 = 0.75.$$

Ahora, C recibe un π -mensaje por ser hijo de B:

$$\pi_C(b_1) = \pi(b_1) = 0,25$$

$$\pi_C(b_2) = \pi(b_2) = 0,75$$

Y actualiza su π -valor:

$$\pi(c_1) = P(c_1/b_1) \pi_C(b_1) + P(c_1/b_2) \pi_C(b_2) = 0.4 \cdot 0.25 + 0.001 \cdot 0.75 = 0.10075$$

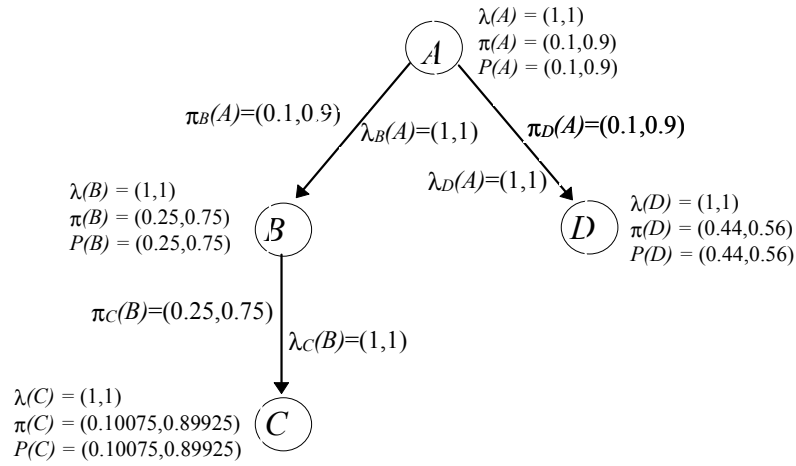
$$\pi(c_2) = P(c_2/b_1) \pi_C(b_1) + P(c_2/b_2) \pi_C(b_2) = 0.89925.$$

A partir de ellos, calculamos las probabilidades de C, multiplicando por los λ -valores y normalizando:

$$P(c_1) = 0.10075$$

$$P(c_2) = 0.89925.$$

El mismo procedimiento se repite para D, y obtenemos el estado inicial S_0 de la red causal:



Estado S_0 de la red.

Supongamos ahora que nos informan de que la esposa ha cenado con otro, es decir, conocemos ahora con certeza que $B = b_1$.

Esta información se irá transmitiendo por la red, haciendo que las probabilidades a priori de los nodos, $P(X)$ cambien a las probabilidades a posteriori, $P^*(X) = P(X/B = b_1)$. En este caso, al ser la evidencia aportada a favor de la hipótesis que queremos probar, lo lógico será que todas estas probabilidades aumenten. En el momento que una variable se actualiza, comienza un flujo de propagación por la red, que en este caso es el siguiente:

- B informa a su padre mediante un λ -mensaje.
- B informa a su hijo mediante un π -mensaje.
- A su vez, A va a informar a su hijo, D, mediante un π -mensaje.

Tras el paso de estos mensajes, todas las variables van a actualizar sus λ y π -valores y sus probabilidades.

Veamos entonces cómo se efectúa la actualización con el algoritmo;

Actualización:

Actualización de B:

- A.1.** Calculamos ahora la probabilidad a posteriori de B , conocido que ha tomado el valor b_1 , que evidentemente será;

$$P^*(b_1) = 1.$$

$$P^*(b_2) = 0.$$

- A.2.** Calculamos $\lambda(B)$;

$$\lambda(b_1) = 1.$$

$$\lambda(b_2) = 0.$$

- A.3.** Enviamos un λ -mensaje al padre de B , A

$$\lambda_B(a_1) = P(b_1/a_1)\lambda(b_1) + P(b_2/a_1)\lambda(b_2) = 0.7 \cdot 1 + 0.3 \cdot 0 = 0.7$$

$$\lambda_B(a_2) = 0.2$$

- A.4.** Enviamos un π -mensaje al hijo de B , C

$$\pi_C(b_1) = 1 \text{ puesto que } B \text{ ha sido instanciada a } b_1.$$

$$\pi_C(b_2) = 0 \text{ puesto que } B \text{ ha sido instanciada a } b_1.$$

Ahora, al haber recibido A y C nuevos mensajes, tienen que actualizar sus valores;

Actualización de C. Al recibir C un π -mensaje, se dispara el procedimiento de actualización C ;

- C.1.** El π -valor de C cambia,

$$\pi(c_1) = P(c_1/b_1)\pi_C(b_1) + P(c_1/b_2)\pi_C(b_2) = 0.4.$$

$$\pi(c_2) = 0.6$$

- C.2.** Calculamos la nueva probabilidad de C

$$P^*(c_1) = 0.4 \quad \alpha = 0.4$$

$$P^*(c_2) = 0.6 \quad \alpha = 0.6$$

- C.3.** No es necesario puesto que C no tiene hijos.

Actualización de A. Al recibir A un λ -mensaje, se dispara el procedimiento de actualización B ;

- B.1.** Actualizamos el λ -valor

$$\lambda(a_1) = \lambda_B(a_1) \lambda_D(a_1) = 0.7$$

$$\lambda(a_2) = \lambda_B(a_2) \lambda_D(a_2) = 0.2$$

B.2. En base al λ -valor, calculamos la probabilidad a posteriori;

$$P^*(a_1) = \alpha \ 0.7 \ 0.1 = 0.07 \ \alpha = 0.28$$

$$P^*(a_2) = \alpha \ 0.2 \ 0.9 = 0.18 \ \alpha = 0.72.$$

Como era de esperar, la probabilidad de que el marido sea infiel ha aumentado, porque la evidencia aportada es a favor de la hipótesis.

B.3. A no tiene padre.

B.4. A envía un π -mensaje a su hijo, D ,

$$\pi_D(a_1) = \pi(a_1) \ \lambda_B(a_1) = 0.1 \ 0.7 = 0.07$$

$$\pi_D(a_2) = \pi(a_2) \ \lambda_B(a_2) = 0.9 \ 0.2 = 0.18.$$

Actualización de D : ahora la variable D , debido a la recepción del π -mensaje, comienza el proceso de actualización C

C.1. El π -valor de D cambia,

$$\pi(d_1) = 0.128$$

$$\pi(d_2) = 0.122$$

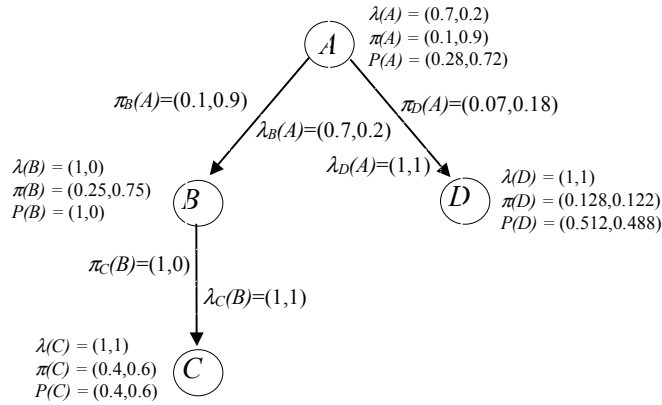
C.2. Calculamos la nueva probabilidad de D

$$P^*(d_1) = 0.512$$

$$P^*(d_2) = 0.488$$

C.3. No es necesario puesto que D no tiene hijos.

Así, tras la instanciación de B a b_1 , la red queda;



Estado S_1 de la red.

Supongamos ahora que tenemos la información de que no se han recibido llamadas telefónicas extrañas en el domicilio, es decir, que sabemos que D ha tomado el valor d_2 .

Nuevamente se iniciará el algoritmo que propagará esta información por la red:

- D enviará un λ -mensaje a su padre, A ,
- A enviará un π -mensaje a su hijo, B .

Pero ahora, al estar B inicializada, el algoritmo se parará ahí, puesto que $P(B) = (1, 0)$, y no podemos permitir que nada cambie ya estos valores. Así, en la ejecución del algoritmo, las variables que ya han sido inicializadas son *extremos muertos*, donde la propagación se para (en el caso de la propagación en árboles).

Hacemos pues el paso A de actualización para la variable D ,

Actualización:

Actualización de D

A.1. Calculamos ahora la probabilidad a posteriori de D ,

$$P^*(d_1) = 0.$$

$$P^*(d_2) = 1.$$

A.2. Calculamos $\lambda(D)$;

$$\lambda(d_1) = 0.$$

$$\lambda(d_2) = 1.$$

A.3. Enviamos un λ -mensaje al padre de D , A

$$\lambda_D(a_1) = P(d_1/a_1)\lambda(d_1) + P(d_2/a_1)\lambda(d_2) = 0.7 \cdot 0 + 0.2 \cdot 1 = 0.2$$

$$\lambda_D(a_2) = 0.6$$

A.4. No se hace puesto que D no tiene hijos.

Actualización de A

B.1. $\lambda(a_1) = \lambda_B(a_1) \lambda_C(a_1) = 0.7 \cdot 0.2 = 0.14$

$$\lambda(a_2) = \lambda_B(a_2) \lambda_C(a_2) = 0.2 \cdot 0.6 = 0.12$$

B.2. $P^*(a_1) = \alpha \cdot 0.014 = 0.1148$

$$P^*(a_2) = \alpha \cdot 0.108 = 0.8852$$

Ahora la probabilidad de a_1 se ha reducido, puesto que la evidencia aportada es en contra de a_1 .

B.3. A no tiene padre.

B.4. Este paso no se realiza pues B está ya instanciado.

Tras estos cálculos se obtiene un estado de la red, S_2 . Este estado es el mismo que obtendríamos si procesásemos la información al revés, es decir, si instanciásemos primero la variable D al valor d_2 , y después la variable B al valor b_1 .