Capítulo 3

Redes bayesianas

En este capítulo vamos a estudiar las redes bayesianas, desde su definición formal (sec. 3.2) hasta los algoritmos de propagación, tanto para poliárboles (sec. 3.3) como para la puerta OR (sec. 3.4). Dado que estas definiciones y algoritmos son difíciles de entender al principio, hasta que el lector se ha familiarizado con ellos, hemos incluido antes un ejemplo médico, que va creciendo en grado de complejidad; su finalidad es mostrar al lector la conexión entre las propiedades formales de las redes bayesianas y el razonamiento de sentido.

3.1 Presentación intuitiva

Antes de presentar formalmente la teoría matemática de las redes bayesianas, intentaremos explicar mediante un ejemplo sencillo, tomado del campo de la medicina, el significado intuitivo de las definiciones y axiomas que luego introduciremos, para mostrar la conexión entre las redes bayesianas y el razonamiento de sentido común. En el ejemplo que vamos a discutir hemos buscado sobre todo una aproximación cualitativa, sin pretender que los factores numéricos sean exactos.

En una red bayesiana, cada nodo corresponde a una variable aleatoria, tal como la edad o el sexo de un paciente, el padecer cierta enfermedad, la presencia de un síntoma o el resultado de una prueba de laboratorio. De aquí en adelante hablaremos indistintamente de nodos y variables, y los denotaremos con letras mayúsculas, tales como X.

Ejemplo 3.1 La red bayesiana no trivial más simple que podemos imaginar consta de dos variables, que llamaremos X e Y_1, y un arco desde la primera a la segunda, como indica la figura 3.1. Por el momento, baste decir que el arco indica generalmente *influencia causal*; más adelante precisaremos el sentido de esta expresión. Utilizaremos frecuentemente el término *enlace* como sinónimo de *arco*.

Por concretar el ejemplo, podemos suponer que X representa Paludismo e Y_1 representa Gota-gruesa, la prueba más habitual para determinar la presencia de dicha enfermedad.

Cuando X es una variable binaria correspondiente a una anomalía, $+x$ indica la presencia de dicha anomalía (en nuestro ejemplo significaría "el paciente tiene paludismo") y $\neg x$ indica su ausencia ("el paciente no tiene paludismo"). Si X representa un test (por ejemplo, Gota-gruesa), $+x$ indica que el test ha dado un resultado positivo y $\neg x$ un resultado negativo.

En la práctica, la información cuantitativa de una red bayesiana viene dada por la probabilidad a priori de los nodos que no tienen padres, $P(x)$, y por la probabilidad condicional

Figura 3.1: Nodo X con un hijo Y_1.

de los nodos con padres, $P(y_1|x)$. Así, en nuestro ejemplo, se supone que conocemos

$$\begin{cases} P(+x) = 0'003 \\ P(\neg x) = 0'997 \end{cases}$$

lo cual significa que el 3 por mil de la población padece paludismo y, por tanto, la probabilidad a priori de que una persona tenga la enfermedad (es decir, la probabilidad cuando no conocemos nada más sobre esa persona) es del 0'3%. En medicina, esta probabilidad a priori se conoce como *prevalencia* de la enfermedad.

También debemos conocer $P(y|x)$, que es la probabilidad condicional del efecto dado el valor de la causa:

$$\left\{ \begin{array}{ll} P(+y_1|+x) = 0'992 & P(+y_1|\neg x) = 0'0006 \\ P(\neg y_1|+x) = 0'008 & P(\neg y_1|\neg x) = 0'9994 \end{array} \right\}$$

El significado de esta probabilidad es el siguiente: cuando hay Paludismo, el test de la Gotagruesa da positivo en el 99'2% de los casos. Este valor se conoce como *sensibilidad* del test. Cuando no hay paludismo, el test da positivo (se dice entonces que ha habido un "falso positivo") en el 0'06% de los casos. La probabilidad de que el test dé negativo cuando la enfermedad buscada está ausente —en nuestro caso es el 99'94%— se llama *especificidad*. En todos los problemas de diagnóstico, no sólo en el campo de la medicina, tratamos de encontrar las pruebas que ofrezcan el grado más alto de sensibilidad y especificidad con el menor coste posible (en términos de dinero, tiempo, riesgo, etc.).[1]

Naturalmente,

$$\begin{cases} P(+y_1|+x) + P(\neg y_1|+x) = 1 \\ P(+y_1|\neg x) + P(\neg y_1|\neg x) = 1 \end{cases}$$

o, en forma abreviada,

$$\sum_{y_1} P(y_1|x) = 1, \quad \forall x \tag{3.1}$$

Conociendo la probabilidad a priori de X y la probabilidad condicional $P(Y_1|X)$, podemos calcular la probabilidad a priori de Y_1 por el teorema de la probabilidad total (ec. (2.9)):

$$\begin{cases} P(+y_1) = P(+y_1|+x) \cdot P(+x) + P(+y_1|\neg x) \cdot P(\neg x) \\ P(\neg y_1) = P(\neg y_1|+x) \cdot P(+x) + P(\neg y_1|\neg x) \cdot P(\neg x) \end{cases}$$

[1] Recordemos que esta definición de *sensibilidad* y *especificidad* es aplicable solamente a un enlace entre variables binarias de tipo presente/ausente o positivo/negativo.

que puede escribirse en forma abreviada como

$$P(y_1) = \sum_x P(y_1|x) \cdot P(x) \tag{3.2}$$

En nuestro ejemplo,

$$\begin{cases} P(+y_1) = 0'00357 \\ P(\neg y_1) = 0'99643 \end{cases}$$

Esto significa que si hacemos el test de la gota gruesa a una persona de la que no tenemos ninguna información, hay un 0'357% de probabilidad de que dé positivo y un 99'643% de que dé negativo.

Vamos a ver ahora cómo podemos calcular la probabilidad a posteriori, es decir, la probabilidad de una variable dada la evidencia observada **e**:

$$P^*(x) \equiv P(x|\mathbf{e}) \tag{3.3}$$

a) Supongamos que la gota gruesa ha dado positivo: $\mathbf{e} = \{+y_1\}$. ¿Cuál es ahora la probabilidad de que nuestro paciente tenga paludismo? Si la prueba tuviera una fiabilidad absoluta, responderíamos que el 100%. Pero como es posible que haya habido un falso positivo, buscamos $P^*(+x)$, es decir, $P(+x|+y_1)$. Para ello, aplicamos el **teorema de Bayes**:

$$P^*(+x) = P(+x|+y_1) = \frac{P(+x) \cdot P(+y_1|+x)}{P(+y_1)} = \frac{0'003 \cdot 0'992}{0'00357} = 0'83263 \tag{3.4}$$

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

También podemos calcular $P^*(\neg x)$:

$$P^*(\neg x) = P(\neg x|+y_1) = \frac{P(\neg x) \cdot P(+y_1|\neg x)}{P(+y_1)} = \frac{0'997 \cdot 0'0006}{0'00357} = 0'16737 \tag{3.5}$$

Esto significa que hay un 16'7% de probabilidad de que haya habido un falso positivo. Naturalmente, se cumple que

$$P^*(+x) + P^*(\neg x) = 1 \tag{3.6}$$

La expresión general del teorema de Bayes que hemos aplicado es

$$P^*(x) = P(x|y) = \frac{P(x) \cdot P(y|x)}{P(y)} \tag{3.7}$$

Por semejanza con el método probabilista clásico (ec. (2.46)), vamos a reescribirla como

$$P^*(x) = \alpha \cdot P(x) \cdot \lambda_{Y_1}(x) \tag{3.8}$$

donde

$$\lambda_{Y_1}(x) \equiv P(\mathbf{e}|x) = P(y_1|x) \tag{3.9}$$

$$\alpha \equiv [P(\mathbf{e})]^{-1} = [P(y_1)]^{-1} \tag{3.10}$$

Vamos a repetir ahora el cálculo anterior aplicando esta reformulación del teorema de Bayes. En primer lugar tenemos que, cuando el test da positivo,

$$\mathbf{e} = \{+y_1\} \implies \begin{cases} \lambda_{Y_1}(+x) = P(+y_1|+x) = 0'992 \\ \lambda_{Y_1}(\neg x) = P(+y_1|\neg x) = 0'0006 \end{cases} \tag{3.11}$$

Esto significa que un resultado positivo en la prueba se explica mucho mejor con la enfermedad presente que con la enfermedad ausente (en la proporción de 0'992/0'0006=1.650), lo cual concuerda, naturalmente, con el sentido común.

Por tanto,

$$\begin{cases} P^*(+x) = \alpha \cdot 0'003 \cdot 0'992 = \alpha \cdot 0'00298 \\ P^*(\neg x) = \alpha \cdot 0'997 \cdot 0'0006 = \alpha \cdot 0'000598 \end{cases}$$

Podríamos calcular α a partir de su definición (3.10), pero resulta mucho más sencillo aplicar la condición de normalización (ec. (3.6)) a la expresión anterior, con lo que se llega a

$$\alpha = [0'00298 + 0'000598]^{-1}$$

y finalmente

$$\begin{cases} P^*(+x) = 0'83263 \\ P^*(\neg x) = 0'16737 \end{cases}$$

que es el mismo resultado que habíamos obtenido anteriormente por la aplicación del teorema de Bayes en su forma clásica. Como en el capítulo anterior, la ecuación (3.8) nos dice que en la probabilidad a posteriori, $P^*(x)$, influyen dos factores: la probabilidad a priori, $P(x)$, y la verosimilitud, $\lambda_{Y_1}(x)$.

b) Y si la gota gruesa diera un resultado negativo, $\mathbf{e} = \{\neg y_1\}$, ¿cuál sería la probabilidad de que el paciente tuviera paludismo? En ese caso,

$$\mathbf{e} = \{\neg y_1\} \implies \begin{cases} \lambda_{Y_1}(+x) = P(\neg y_1|+x) = 0'008 \\ \lambda_{Y_1}(\neg x) = P(\neg y_1|\neg x) = 0'9994 \end{cases} \tag{3.12}$$

Es decir, un resultado negativo en la prueba de la gota gruesa se explica mucho mejor (en la proporción de 0'9994/0'008 = 125) cuando no hay paludismo que cuando lo hay; en otras palabras, para $\neg y_1$, el valor $\neg x$ es 125 veces más verosímil que $+x$.

Aplicando la ecuación (3.8) como en el caso anterior, obtenemos

$$\begin{cases} P^*(+x) = \alpha \cdot 0'003 \cdot 0'008 = 0'000024 \\ P^*(\neg x) = \alpha \cdot 0'997 \cdot 0'9994 = 0'999976 \end{cases}$$

donde hemos calculado α por normalización.

El resultado tan bajo para $P^*(+x)$ se explica por dos razones: por un lado, la probabilidad a priori era de sólo un 0'3%; por otro, la alta especificidad de la prueba (99'94%) es un argumento convincente para descartar la enfermedad.

De nuevo comprobamos que en la probabilidad a posteriori influyen la probabilidad a priori y la verosimilitud. □

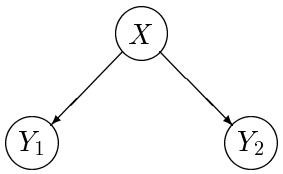

Figura 3.2: Nodo X con dos hijos.

Ejemplo 3.2 Vamos a ampliar el modelo anterior añadiendo un nuevo efecto producido por el paludismo, la Fiebre, que representaremos mediante la variable Y_2, tal como muestra la figura 3.2.

La probabilidad condicional para este segundo enlace XY_2 viene definida por

$$\left\{\begin{array}{ll} P(+y_2|+x) = 0'98 & P(+y_2|\neg x) = 0'017 \\ P(\neg y_2|+x) = 0'02 & P(\neg y_2|\neg x) = 0'983 \end{array}\right\}$$

que indica la probabilidad de que un paciente (o una persona, en general) tenga fiebre dependiendo de si tiene o no paludismo. Vemos aquí que, para el paludismo, la fiebre tiene mucha menor especificidad (98'3%) que la gota gruesa (99'94%). Así, este sencillo modelo tiene en cuenta que hay muchas otras causas que pueden producir fiebre, aunque no las incluya explícitamente.

Aplicando el teorema de la probabilidad total (ec. (2.9)) podemos calcular la probabilidad a priori de que un enfermo tenga fiebre,

$$P(+y_2) = \sum_x P(+y_2|x) \cdot P(x) = 0'01989$$

pero éste es un resultado que carece de importancia para el diagnóstico.

a) Supongamos que encontramos un paciente con fiebre, $\mathbf{e} = \{+y_2\}$, y queremos hallar la probabilidad de que tenga paludismo. En primer lugar, expresamos el teorema de Bayes en forma normalizada:

$$P^*(x) = \alpha \cdot P(x) \cdot \lambda_{Y_2}(x) \tag{3.13}$$

Ahora α vale $[P(+y_2)]^{-1}$, pero podemos prescindir de su significado y tratarla simplemente como una constante de normalización.

Para un paciente con fiebre,

$$\mathbf{e} = \{+y_2\} \implies \left\{\begin{array}{l} \lambda_{Y_2}(+x) = P(+y_2|+x) = 0'98 \\ \lambda_{Y_2}(\neg x) = P(+y_2|\neg x) = 0'017 \end{array}\right. \tag{3.14}$$

de modo que

$$\left\{\begin{array}{l} P^*(+x) = \alpha \cdot 0'003 \cdot 0'98 = 0'148 \\ P^*(\neg x) = \alpha \cdot 0'997 \cdot 0'017 = 0'852 \end{array}\right.$$

lo cual significa que hay un 14'8% de probabilidad de que el paciente tenga paludismo. Compárese con el 83'3% correspondiente a un resultado positivo de la gota gruesa (ec. (3.4)). La diferencia se debe de que esta prueba es un signo muy específico de la enfermedad, mientras que la fiebre puede estar producida por muchas otras causas.

b) Vamos a estudiar ahora el caso en que tenemos las dos observaciones y ambas indican la presencia de la enfermedad: $\mathbf{e} = \{+y_1, +y_2\}$. Al intentar calcular la probabilidad de que esa persona tenga paludismo, $P(+x|+y_1, +y_2)$ nos damos cuenta de que nos falta información, pues para aplicar el teorema de Bayes,

$$P(x|+y_1, +y_2) = \frac{P(+y_1, +y_2|\, x) \cdot P(x)}{P(+y_1, +y_2)} \tag{3.15}$$

necesitamos conocer $P(+y_1, +y_2|\, x)$ y $P(+y_1, +y_2)$.

Con la información disponible es imposible calcular estas expresiones. Por ello vamos a introducir la *hipótesis de independencia condicional*. Examinemos primero el caso en que sabemos con certeza que hay paludismo ($X = +x$). Entonces es razonable pensar que la probabilidad de que el paciente tenga o no tenga fiebre no depende de si hemos realizado el test de la gota gruesa ni del resultado que éste haya dado: la fiebre depende sólo de si hay paludismo (dando por supuesto, como parece razonable, que las demás causas de fiebre no influyen en el resultado del test). La afirmación "conociendo $X = x$, el valor de y_2 no depende del de y_1" se expresa matemáticamente como

$$P(y_2|+x, y_1) = P(y_2|+x) \tag{3.16}$$

o dicho de otro modo[2]

$$P(y_1, y_2|+x) = P(y_1|+x) \cdot P(y_2|+x) \tag{3.17}$$

Observar que estas expresiones son simétricas para Y_1 e Y_2.

Supongamos ahora que *no* hay paludismo ($X = \neg x$). La probabilidad de que el paciente presente fiebre no depende de si la gota gruesa ha dado negativo (como era de esperar) o ha dado un falso positivo por alguna extraña razón. Así tenemos

$$P(y_2|\neg x, y_1) = P(y_2|\neg x) \tag{3.18}$$

o lo que es lo mismo

$$P(y_1, y_2|\neg x) = P(y_1|\neg x) \cdot P(y_2|\neg x) \tag{3.19}$$

Uniendo las ecuaciones (3.17) y (3.19), tenemos

$$P(y_1, y_2|x) = P(y_1|x) \cdot P(y_2|x) \tag{3.20}$$

que es lo que se conoce como *independencia condicional*. Definiendo

$$\lambda(x) \equiv P(y_1, y_2|x) \tag{3.21}$$

podemos expresar dicha propiedad como

$$\lambda(x) = \lambda_{Y_1}(x) \cdot \lambda_{Y_2}(x) \tag{3.22}$$

Con esta hipótesis ya podemos calcular la probabilidad buscada. La ecuación (3.15) es equivalente a

$$P^*(x) = \alpha \cdot P(x) \cdot \lambda(x) \tag{3.23}$$

[2]Ambas expresiones son equivalentes cuando $P(+x) \neq 0$, pues

$$P(y_1, y_2|+x) = \frac{P(y_1, y_2, +x)}{P(+x)} = \frac{P(y_2|y_1, +x) \cdot P(y_1, +x)}{P(+x)} = P(y_2|+x, y_1) \cdot P(y_1|+x)$$

En nuestro ejemplo, a partir de las ecuaciones (3.11) y (3.14) tenemos

$$\mathbf{e} = \{+y_1, +y_2\} \implies \begin{cases} \lambda(+x) = 0'97216 \\ \lambda(-x) = 0'0000102 \end{cases} \tag{3.24}$$

El valor de α se calcula al normalizar, obteniendo así

$$\begin{cases} P^*(+x) = 0'99653 \\ P^*(\neg x) = 0'00347 \end{cases}$$

Naturalmente, cuando hay dos hallazgos a favor del paludismo, la probabilidad resultante (99'7%) es mucho mayor que la correspondiente a cada uno de ellos por separado (83'3% y 14'8%).

En realidad, lo que hemos hecho en este apartado no es más que aplicar el método probabilista clásico en forma normalizada (sec. 2.4); puede comprobarlo comparando las ecuaciones (3.21) y (3.22) con la (2.47) y la (2.48), respectivamente.

c) En el caso de que tengamos un hallazgo a favor y otro en contra, podemos ponderar su influencia mediante estas mismas expresiones. Por ejemplo, si hay fiebre ($+y_2$) pero hay un resultado negativo en la prueba de la gota gruesa ($\neg y_1$), las ecuaciones (3.12), (3.14) y (3.22) nos dicen que

$$\mathbf{e} = \{\neg y_1, +y_2\} \implies \begin{cases} \lambda(+x) = 0'008 \quad \cdot 0'98 \quad = 0'00784 \\ \lambda(\neg x) = 0'9994 \cdot 0'017 = 0'01699 \end{cases} \tag{3.25}$$

Vemos que hay más evidencia a favor de $\neg x$ que de $+x$ (en la proporción aproximada de 2 a 1), debido sobre todo al 0'008 correspondiente a la gota gruesa, lo cual es un reflejo de la alta sensibilidad de esta prueba (99'2%). Es decir, si hubiera paludismo, es casi seguro que lo habríamos detectado; al no haberlo detectado, tenemos una buena razón para descartarlo.

Al tener en cuenta además la probabilidad a priori de la enfermedad, nos queda finalmente

$$\begin{cases} P^*(+x) = \alpha \cdot 0'003 \cdot 0'00784 = 0'0014 \\ P^*(\neg x) = \alpha \cdot 0'997 \cdot 0'01699 = 0'9986 \end{cases}$$

Por tanto, la ponderación de la evidencia ha modificado la probabilidad desde 0'3% (valor a priori) hasta 0'14% (valor a posteriori).

De nuevo hemos aplicado el método probabilista clásico en forma normalizada.

d) Aún podemos obtener más información de este ejemplo. Imaginemos que tenemos un paciente con fiebre ($Y_2 = +y_2$) y todavía no hemos realizado la prueba de la gota gruesa. ¿Qué probabilidad hay de que ésta dé un resultado positivo o negativo? Es decir, ¿cuánto vale $P(y_1|+y_2)$?

Por la teoría elemental de la probabilidad sabemos que

$$\begin{aligned} P^*(y_1) = P(y_1|+y_2) &= \sum_x P(y_1|\,x, +y_2) \cdot P(x|+y_2) \\ &= \sum_x P(y_1|\,x, +y_2) \cdot \frac{P(x, +y_2)}{P(+y_2)} \end{aligned}$$

Aplicando la independencia condicional dada en (3.17) y definiendo[3]

$$\pi_{Y_1}(x) \equiv P(x, +y_2) = P(x) \cdot P(+y_2 | \, x) \tag{3.26}$$

$$\alpha \equiv [P(+y_2)]^{-1} \tag{3.27}$$

podemos reescribir la expresión anterior como

$$P^*(y_1) = \alpha \cdot \sum_x P(y_1|x) \cdot \pi_{Y_1}(x) \tag{3.28}$$

Sustituyendo los valores numéricos, tenemos que

$$\mathbf{e} = \{+y_2\} \implies \begin{cases} \pi_{Y_1}(+x) = 0'003 \cdot 0'98 & = 0'00294 \\ \pi_{Y_1}(\neg x) = 0'997 \cdot 0'017 & = 0'01695 \end{cases} \tag{3.29}$$

y finalmente

$$\begin{cases} P^*(+y_1) = 0'14715 \\ P^*(\neg y_1) = 0'85285 \end{cases} \tag{3.30}$$

Resulta muy interesante comparar la ecuación (3.28) con (3.2). Al buscar la probabilidad a priori $P(y_1)$ utilizábamos $P(x)$; ahora, al calcular $P^*(y_1)$, utilizamos $\pi_{Y_1}(x)$, que indica la probabilidad de X tras considerar la evidencia relativa a X *diferente* de Y_1.

Vemos así cómo 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 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 Y_1 puede ser fuente de información u objeto de predicción, dependiendo de cuáles sean los hallazgos disponibles y el objetivo del diagnóstico. □

Ejemplo 3.3 Consideremos una red bayesiana en que un nodo X, que representa la variable Paludismo, tiene dos padres, U_1 = País-de-origen y U_2 = Tipo-sanguíneo, dos de los factores que influyen en la probabilidad de contraer la enfermedad, tal como muestra la figura 3.3.

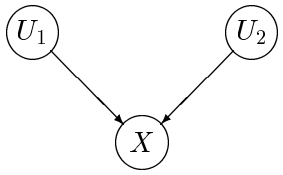

Figura 3.3: Nodo X con dos padres.

La variable U_1 podría tener muchos valores, tantos como países de origen quisiéramos considerar. Sin embargo, vamos a suponer que los agrupamos en tres zonas, de alto, medio

[3]Puede resultar extraño al lector que $\pi_{Y_1}(x)$ lleve el subíndice Y_1 a pesar de que depende del valor de la variable Y_2. El motivo es que $\pi_{Y_1}(x)$ porque recoge toda la evidencia relativa a X **diferente de** Y_1. Daremos una definición más precisa en la sección 3.3.1.

y bajo riesgo, que denotaremos por u_1^+, u_1^0 y u_1^-, respectivamente. La variable U_2 (Tipo-sanguíneo) puede tomar dos valores: $u_2^\dagger$ ó $u_2^\ddagger$.

Las probabilidades a priori para U_1 y U_2 son:

$$\begin{cases} P(u_1^+) = 0'10 \\ P(u_1^0) = 0'10 \\ P(u_1^-) = 0'80 \end{cases} \qquad \begin{cases} P(u_2^\dagger) = 0'60 \\ P(u_2^\ddagger) = 0'40 \end{cases} \tag{3.31}$$

Esto significa que la mayor parte de las personas que vamos a examinar proceden de una zona de bajo riesgo, u_1^-, y que el primer tipo sanguíneo, $u_2^\dagger$, es más frecuente que el segundo.

Las probabilidades condicionadas aparecen en la tabla 3.1. En ella vemos que, efectivamente, la zona u_1^+ es la de mayor riesgo y u_1^- la de menor. También observamos que el tipo sanguíneo $u_2^\dagger$ posee mayor inmunidad que el $u_2^\ddagger$.

$U_2 \setminus U_1$	u_1^+	u_1^0	u_1^-
$u_2^\dagger$	0'015	0'003	0'0003
$u_2^\ddagger$	0'022	0'012	0'0008

Tabla 3.1: Probabilidad de padecer paludismo, $P(+x|u_1, u_2)$.

La probabilidad de que una persona (de la que no tenemos ninguna información) padezca paludismo es

$$P(x) = \sum_{u_1,u_2} P(x|u_1, u_2) \cdot P(u_1, u_2) \tag{3.32}$$

De nuevo tenemos el problema de que no conocemos $P(u_1, u_2)$. Podemos entonces hacer la hipótesis de *independencia a priori* entre ambas variables; es decir, suponemos que los tipos sanguíneos se distribuyen por igual en las tres zonas de riesgo. Ésta es una hipótesis que habría que comprobar empíricamente. Si llegáramos a la conclusión de que existe una correlación entre ambas variables, deberíamos trazar un arco desde la una hasta la otra e introducir las correspondientes tablas de probabilidades condicionadas.

Estamos observando aquí una propiedad esencial de las RR.BB.: no sólo los arcos aportan información sobre dependencias causales, sino que también *la ausencia de un arco es una forma (implícita) de aportar información.* En nuestro caso implica que U_1 y U_2 son independientes. Matemáticamente se expresa así:

$$P(u_2|u_1) = P(u_2) \tag{3.33}$$

o bien

$$P(u_1, u_2) = P(u_1) \cdot P(u_2) \tag{3.34}$$

Con esta hipótesis podemos por fin calcular la probabilidad de X:

$$P(x) = \sum_{u_1} \sum_{u_2} P(x|u_1, u_2) \cdot P(u_1) \cdot P(u_2) \tag{3.35}$$

En nuestro caso, el valor obtenido es $P(+x) = 0'003$, que concuerda con el de los ejemplos anteriores.

a) Supongamos que nos enteramos de que la persona en cuestión procede de una zona de alto riesgo. ¿Cual es la probabilidad de que padezca la enfermedad? Una de las formas posibles de realizar el cálculo es ésta:

$$P^*(x) = P(x|u_1^+) = \frac{P(x, u_1^+)}{P(u_1^+)}$$

Si definimos

$$\pi(x) \equiv P(x, \mathbf{e}) = P(x, u_1^+) \tag{3.36}$$

$$\alpha \equiv [P(\mathbf{e})]^{-1} = [P(u_1^+)]^{-1} \tag{3.37}$$

la ecuación anterior se convierte en

$$P^*(x) = \alpha \cdot \pi(x) \tag{3.38}$$

Podemos obtener $\pi(x)$ del siguiente modo:

$$\pi(x) = \sum_{u_2} P(x, u_1^+, u_2) = \sum_{u_2} P(x|u_1^+, u_2) \cdot P(u_1^+, u_2)$$

y aplicando la independencia a priori de las causas podemos expresar la ecuación anterior como

$$\pi(x) = \sum_{u_1} \sum_{u_2} P(x|u_1, u_2) \cdot \pi_X(u_1) \cdot \pi_X(u_2) \tag{3.39}$$

que es un resultado completamente general.

En el ejemplo que estamos tratando,

$$\mathbf{e} = \{u_1^+\} \implies \left\{ \begin{array}{ll} \pi_X(u_1^+) = P(u_1^+) & \\ \pi_X(u_1^0) = 0 & \pi_X(u_2^\dagger) = P(u_2^\dagger) \\ \pi_X(u_1^-) = 0 & \pi_X(u_2^\ddagger) = P(u_2^\ddagger) \end{array} \right\} \tag{3.40}$$

y en consecuencia

$$\mathbf{e} = \{u_1^+\} \implies \left\{ \begin{array}{l} \pi(+x) = 0'00194 \\ \pi(\neg x) = 0'09806 \end{array} \right. \tag{3.41}$$

Sustituyendo este resultado en la ecuación (3.38) y normalizando (en este caso, $\alpha = 10$), hallamos que $P^*(+x) = 0'0194$; es decir, una persona originaria de una zona de alto riesgo tiene una probabilidad del 2% de padecer paludismo (frente al 0'3% general).

Las expresiones $\pi_X(u_i)$ que hemos utilizado en la deducción no son nuevas: aparecieron ya en la ecuación (3.26). Recordemos que el significado de $\pi_X(u_i)$ es que transmite a X el impacto de toda la evidencia relativa a U_i. Como no hay evidencia relativa a U_2, $\pi_X(u_2)$ coincide con la probabilidad a priori.

b) Imaginemos ahora que por alguna razón tenemos certeza absoluta de que el enfermo padece paludismo. Antes de hacer un análisis de sangre, podemos predecir qué resultado es más probable, considerando cuál de los dos tipos sanguíneos explica mejor la presencia de la enfermedad:

$$P^*(u_2) = P(u_2|+x) = \frac{P(u_2) \cdot P(+x|u_2)}{P(+x)}$$

o bien

$$P^*(u_2) = \alpha \cdot P(u_2) \cdot \lambda_X(u_2) \tag{3.42}$$

donde

$$\lambda_X(u_2) \equiv P(+x|u_2) = \sum_{u_1} P(+x|u_1, u_2) \cdot P(u_1) \tag{3.43}$$

que en nuestro ejemplo vale

$$\mathbf{e} = \{+x\} \implies \begin{cases} \lambda_X(u_2^\dagger) = 0'00204 \\ \lambda_X(u_2^\ddagger) = 0'00444 \end{cases} \tag{3.44}$$

Efectivamente, los valores de la tabla 3.1 nos han llevado a la conclusión de que el paludismo se explica mejor con el tipo sanguíneo $u_2^\ddagger$. Aplicando (3.42), obtenemos

$$\begin{cases} P^*(u_2^\dagger) = 0'408 \\ P^*(u_2^\ddagger) = 0'592 \end{cases} \tag{3.45}$$

Observamos que inicialmente el tipo $u_2^\dagger$ era el más probable (60%), pero ahora es el menos probable (40'8%) porque explica peor el paludismo.

El cálculo que hemos realizado para X y U_2 es idéntico al que hicimos en el ejemplo 3.1.a para Y_1 y X. Vemos de nuevo que un mismo nodo puede ser fuente de información u objeto de predicción, dependiendo de la evidencia disponible.

c) Mostraremos ahora otra de las propiedades más características de las RR.BB.: la aparición de correlaciones entre los padres de un nodo. Continuando con el caso anterior, supongamos que además de tener la certeza de que el enfermo padece paludismo sabemos que procede de un país de alto riesgo; es decir, $\mathbf{e} = \{+x, u_1^+\}$. Aplicaremos de nuevo la ecuación (3.42), aunque ahora

$$\lambda_X(u_2) \equiv P(+x, u_1^+|u_2) = P(+x|u_1^+, u_2) \cdot P(u_1^+|u_2)$$

La independencia condicional nos dice que $P(u_1^+|u_2) = P(u_1^+)$, y tenemos, por tanto,

$$\begin{aligned} \lambda_X(u_2) &= P(x|u_1^+, u_2) \cdot P(u_1^+) \\ &= \sum_{u_1} P(x|u_1, u_2) \cdot \pi_X(u_1) \end{aligned} \tag{3.46}$$

donde $\pi_X(u_1)$ es el vector que apareció en la ecuación (3.40). Al realizar los cálculos obtenemos

$$\mathbf{e} = \{+x, u_1^+\} \implies \begin{cases} \lambda_X(u_2^\dagger) = 0'0015 \\ \lambda_X(u_2^\ddagger) = 0'0026 \end{cases} \tag{3.47}$$

y de ahí

$$\begin{cases} P^*(u_2^\dagger) = 0'464 \\ P^*(u_2^\ddagger) = 0'536 \end{cases} \tag{3.48}$$

Si comparamos este resultado con el de la ecuación (3.45), observamos que la probabilidad de $u_2^\dagger$ ha aumentado del 40'8% al 46'4% como resultado de conocer la zona de origen: $U_1 = u_1^+$. Éste es el fenómeno que queríamos mostrar. *A priori*, es decir, antes de conocer el valor de x,

U_1 y U_2 eran *independientes*, por lo que la probabilidad de u_2 no variaba al conocer el valor u_1 (cf. ec. (3.33)). Sin embargo, la independencia se pierde tanto al conocer "$X = +x$" como "$X = \neg x$". Dicho de otro modo,

$$P(u_2|u_1) = P(u_2) \tag{3.49}$$

$$P(u_2|u_1, x) \neq P(u_2|x) \tag{3.50}$$

Recordando el ejemplo 3.2, vemos que allí ocurría precisamente lo contrario: las variables Y_1 e Y_2 estaban correlacionadas a priori, pero se volvían *condicionalmente independientes* al conocer el valor de X. Esta asimetría en las relaciones de independencia es un reflejo del sentido de la causalidad, es decir, de la diferencia entre causas y efectos. □

Ejemplo 3.4 Por último, consideremos el caso en que tenemos un nodo con dos causas y dos efectos (fig. 3.4). Las probabilidades condicionadas son las mismas que en los ejemplos anteriores. Por no extender demasiado esta sección, vamos a considerar solamente el caso en que tenemos un paciente que procede de una zona de alto riesgo y presenta fiebre, pero la prueba de la gota gruesa ha dado un resultado negativo. Es decir, $\mathbf{e} = \{u_1^+, \neg y_1, +y_2\}$.

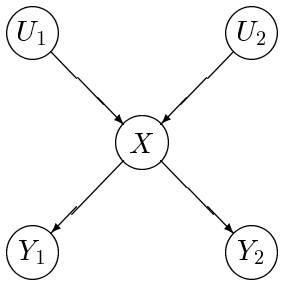

Figura 3.4: Nodo X con dos padres y dos hijos.

El teorema de Bayes nos dice que

$$P^*(x) = P(x|u_1^+, \neg y_1, +y_2) = \frac{P(x) \cdot P(u_1^+, \neg y_1, +y_2|x)}{P(u_1^+, \neg y_1, +y_2)} \tag{3.51}$$

Nuevamente necesitamos utilizar unos valores, $P(u_1^+, \neg y_1, +y_2|x)$, que no conocemos. (Si tuviéramos estos valores podríamos calcular también el denominador de la fracción.) Hemos introducido ya dos hipótesis:

1. Independencia a priori de los nodos que no tienen ningún antepasado común.
2. Independencia condicional de los dos efectos de X cuando conocemos con certeza el valor de X.

Vamos a enunciar ahora la tercera y última hipótesis, la independencia condicional (para cada valor x) entre los padres y los hijos de X:

$$P(y_1, y_2|x, u_1, u_2) = P(y_1, y_2|x) \tag{3.52}$$

o, lo que es lo mismo,

$$P(u_1, u_2, y_1, y_2|x) = P(u_1, u_2|x) \cdot P(y_1, y_2|x) \tag{3.53}$$

La interpretación de estas dos ecuaciones es clara: la probabilidad de los efectos de X depende solamente del valor que toma X, no de la combinación de factores que nos ha llevado a dicho valor. En nuestro ejemplo significa que, si hay certeza de que una persona padece paludismo, la probabilidad de que tenga fiebre y de que detectemos la enfermedad en la prueba de laboratorio no depende del país de origen ni del tipo sanguíneo. Lo mismo podemos decir de la ausencia de paludismo.[4]

De la ecuación (3.53) se deduce fácilmente, sumando sobre u_2, que

$$P(u_1, y_1, y_2|x) = P(u_1|x) \cdot P(y_1, y_2|x) \tag{3.54}$$

con lo que la ecuación (3.51) se convierte en

$$P^*(x) = \alpha \cdot \pi(x) \cdot \lambda(x) \tag{3.55}$$

Recordemos que ya habíamos definido anteriormente $\pi(x)$ y $\lambda(x)$:

$$\pi(x) \equiv P(x) \cdot P(u_1^+|x) = P(x, u_1^+) \tag{3.56}$$
$$\lambda(x) \equiv P(\neg y_1, +y_2|x) \tag{3.57}$$

y que sus valores estaban dados por (3.41) y (3.25), respectivamente. Tras unos cálculos sencillos obtenemos que $P^*(+x) = 0'0090$; es decir, con estos hallazgos, la probabilidad de que el paciente tenga paludismo es menor del 1%.

Podríamos calcular ahora la probabilidad del tipo sanguíneo en función de la evidencia, $P^*(u_2)$, pero lo vamos a omitir para no alargar más la exposición.

La ecuación (3.55) es la fórmula fundamental para el cálculo de la probabilidad en redes bayesianas. En ella aparecen dos términos importantes, $\pi(x)$ y $\lambda(x)$. El primero de ellos transmite el impacto de la evidencia correspondiente a **las causas** de X. En nuestro caso, el único hallazgo "por encima" de X era $U_1 = u_1^+$. Si no tuviéramos ninguna evidencia, $\pi(x)$ sería simplemente la probabilidad a priori $P(x)$.

El segundo, $\lambda(x)$, transmite el impacto de la evidencia correspondiente a **los efectos** de X. En el ejemplo anterior, recogía la influencia de $\neg y_1$ y $+y_2$. Si no tuviéramos ninguna evidencia, $\lambda(x)$ sería un vector constante y podríamos prescindir de él al aplicar la ecuación (3.55), sin alterar el resultado.

De las tres propiedades de independencia anteriores —ecs. (3.20), (3.34) y (3.53)— que no son más que la manifestación de la **separación direccional** (sec. 3.2.2) para esta pequeña red, se deduce que

$$P(y_1, y_2, x, u_1, u_2) = P(y_1|x) \cdot P(y_2|x) \cdot P(x|u_1, u_2) \cdot P(u_1) \cdot P(u_2) \tag{3.58}$$

Esta expresión se conoce como *factorización de la probabilidad* en una red bayesiana (cf. teorema 3.7, pág. 52). □

[4] Si por alguna razón pensáramos que esta hipótesis no es cierta, deberíamos añadir a nuestro modelo nuevos arcos con el fin de representar las influencias existentes (por ejemplo, entre el país de origen y otras causas de la fiebre) y asignarles las tablas de probabilidad oportunas.

Recapitulación

En esta sección hemos visto las propiedades más importantes de las RR.BB. En primer lugar, que la red contiene información cualitativa (la estructura del grafo) y cuantitativa (las probabilidades a priori y condicionales). Esta red constituye nuestro modelo causal y —salvo que introduzcamos algún mecanismo de aprendizaje— es invariable.

El proceso de diagnóstico consiste en introducir la evidencia disponible (asignar valores a las variables conocidas) y calcular la probabilidad a posteriori de las variables desconocidas. Se trata en realidad de un proceso de inferencia, aunque no es simbólica sino numérica.

Hemos visto además que este modelo permite tanto un razonamiento diagnóstico (cuál es la causa más probable) como predictivo (qué valor de cierta variable aparecerá con mayor probabilidad). Por otra parte, hemos comentado ya que una ventaja de las RR.BB. es que un mismo nodo puede ser fuente de información u objeto de predicción dependiendo de cuál sea la evidencia disponible, como ocurría con X o con Y_1 en los ejemplos anteriores.

Y hemos comprobado también en el ejemplo 4 que es posible realizar un cálculo incremental, modificando la probabilidad de las variables a medida que va llegando nueva evidencia, sin tener que recalcular todos los mensajes $\pi()$ y $\lambda()$.

3.2 Definición formal de red bayesiana

En la sección anterior hemos presentado de forma intuitiva qué son las redes bayesianas y cómo se propaga la evidencia, insistiendo en la importancia de las hipótesis de independencia. Ahora vamos a dar una definición matemática formal.

3.2.1 Estructura de la red. Teoría de grafos

Nuestro punto de partida consiste en un conjunto finito de *nodos* $\bar{X}$. Cada uno de ellos representa una *variable*, que puede ser discreta o continua (aunque en este texto sólo vamos a manejar variables discretas). Esta relación biunívoca entre nodos y variables nos permite emplear indistintamente ambos términos. Como vimos en el capítulo anterior, los valores de una variable deben constituir un conjunto exclusivo y exhaustivo.

Sin embargo, una diferencia importante respecto del método probabilista clásico (sec. 2.4) es que las redes bayesianas no necesitan suponer que los diagnósticos son exclusivos y exhaustivos, y por tanto no es necesario tener una variable D que represente todos los posibles diagnósticos; por ejemplo, en vez de una variable llamada D=Enfermedad, cuyos valores representasen los posibles diagnósticos correspondientes a la fiebre: neumonía, amigdalitis, paludismo, etc., en la red bayesiana tendríamos una variable Neumonía —que puede tomar dos valores (neumonía-presente y neumonía-ausente) o más de dos valores (neumonía-ausente, neumonía-leve, neumonía-moderada y neumonía-severa), dependiendo del grado de precisión que necesitemos en el diagnóstico—, otra variable Amigdalitis, Paludismo, etc. De este modo, la red bayesiana puede ofrecer dos o más diagnósticos a la vez (por ejemplo, amigdalitis-severa y neumonía-leve), lo cual era imposible con el método probabilista clásico.[5]

Introducimos a continuación algunas definiciones básicas sobre grafos:

[5]Las redes de semejanza de Heckerman [27] consituyen una notable excepción, pues en cada una de ellas hay un nodo principal, que representa los diagnósticos (supuestamente exclusivos y exhaustivos). Aunque en esto coinciden con el método probabilista clásico, se diferencian de él en que permiten que el nodo principal tenga padres y que los hijos puedan tener hijos a su vez, e incluso que haya bucles en la red.

▷ **Arco.** Es un par ordenado de nodos (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, tal como muestran las figuras de los ejemplos anteriores.

▷ **Grafo dirigido.** Es un par $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ donde $\mathcal{N}$ es un conjunto de nodos y $\mathcal{A}$ un conjunto de arcos definidos sobre los nodos.

Si hubiéramos definido los arcos como pares no ordenados, tendríamos un grafo no dirigido.[6]

En el contexto de los **grafos dirigidos**, tenemos las siguientes definiciones:

▷ **Padre.** X es un *padre* de Y si y sólo si existe un arco (X, Y).

Los padres de X se representan como $pa(X)$. Por semejanza con el convenio utilizado para variables y sus valores, $pa(x)$ representará el vector formado al asignar un valor a cada nodo del conjunto $pa(X)$.

▷ **Hijo.** Y es un *hijo* de X si y sólo si existe un arco (X, Y).

▷ **Antepasado.** X es un *antepasado* de Z si y sólo si existe (al menos) un nodo Y tal que X es padre de Y e Y es antepasado de Z.

▷ **Descendiente.** Z es un *descendiente* de X si y sólo si X es un antepasado de Z.

▷ **Familia X.** Es el conjunto formado por X y los padres de X, $pa(X)$.

▷ **Nodo terminal.** Es el nodo que no tiene hijos.

Ejemplo 3.5 En la figura 3.5, los padres de D son A y B: $pa(D) = \{A, B\}$. Los hijos de D son G y H. Los antepasados de G son A, B y D. Los descendientes de A son D, G y H. Las nueve familias (tantas como nodos) son $\{A\}$, $\{B\}$, $\{C\}$, $\{D, A, B\}$, $\{E, C\}$, $\{F, C\}$, $\{G, D\}$, $\{H, D, E\}$ e $\{I, E\}$.

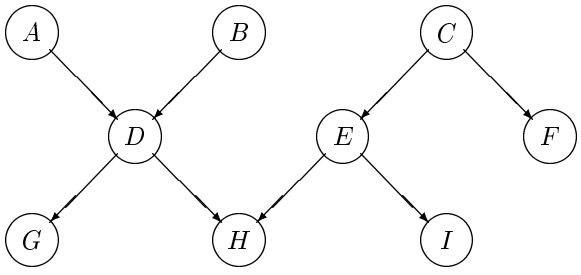

Figura 3.5: Un pequeño poliárbol.

[6] Las redes de Markov se basan en grafos no dirigidos, mientras que las redes bayesianas corresponden a grafos dirigidos.

▷ **Camino.** Un *camino* entre X_1 y X_N en una sucesión de nodos $\{X_1, \ldots, X_N\}$ pertenecientes a un grafo $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, tal que $X_i \neq X_j$ para $1 \leq i < j \leq N$ y

$$(X_i, X_{i+1}) \in \mathcal{A} \quad \text{ó} \quad (X_{i+1}, X_i) \in \mathcal{A}, \qquad \forall i, 1 \leq i < N$$

Es decir, dos nodos consecutivos de un camino —X_i y X_{i+1}— están unidos por un arco del primero al segundo o viceversa. Observe que esta definición corresponde a lo que en otros lugares se conoce como *camino abierto*.

▷ **Ciclo.** Es una sucesión de nodos $\{X_1, \ldots, X_N\}$ pertenecientes a un grafo $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, tal que *(1)* $X_i \neq X_j$ para $1 \leq i < j \leq N$, *(2)* para todo $i < N$ existe en $\mathcal{A}$ un arco (X_i, X_{i+1}), y *(3)* existe además un arco (X_N, X_1).

▷ **Bucle.** Sucesión de nodos $\{X_1, \ldots, X_N\}$ pertenecientes a un grafo $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, tal que *(1)* $X_i \neq X_j$ para $1 \leq i < j \leq N$, *(2)* para todo $i < N$ existe en $\mathcal{A}$ un arco (X_i, X_{i+1}) ó (X_{i+1}, X_i), *(3)* existe además un arco (X_N, X_1) ó (X_1, X_N) y *(4)* los arcos no forman un ciclo.

▷ **Grafo acíclico.** Es el grafo en que no hay ciclos.

Tanto el ciclo como el bucle corresponden a lo que a veces se denominan *caminos cerrados simples*. La diferencia es que en un ciclo los arcos van de cada nodo al siguiente (nunca a la inversa), mientras que la definición de bucle permite que los arcos tengan cualquiera de los dos sentidos, con la única condición de que no formen un ciclo. La distinción entre ambos es muy importante para el tema que nos ocupa, pues las redes bayesianas se definen a partir de los **grafos dirigidos acíclicos**, lo cual permite que contengan bucles pero no que contengan ciclos.

Ejemplo 3.6 En la figura 3.6.*a*, vemos que entre B y C hay dos *caminos*: $\{B, A, C\}$ y $\{B, D, C\}$, y lo mismo ocurre en 3.6.*b* y 3.6.*c*. El primero de estos tres grafos es un *ciclo*, mientras que que los dos últimos son *bucles*. Por eso estos dos últimos podrían servir para definir redes bayesianas, pero el primero no. □

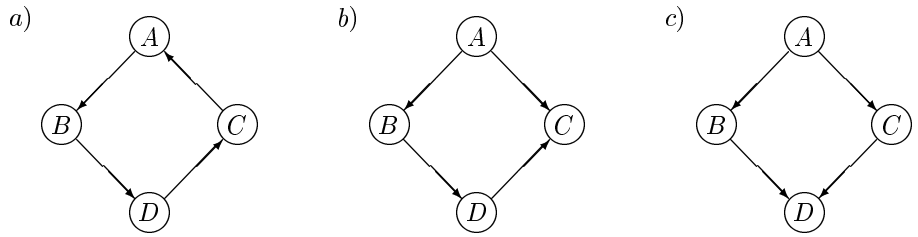

Figura 3.6: Un ciclo y dos bucles.

▷ **Grafo conexo.** Un grafo es *conexo* si entre dos cualesquiera de sus nodos hay al menos un camino.

Por tanto, un grafo no conexo es aquél que está formado por dos o más partes inconexas entre sí. Todo grafo conexo ha de pertenecer a una de las dos categorías siguientes:

▷ **Grafo simplemente conexo o poliárbol.** Un grafo es *simplemente conexo* si entre dos cualesquiera de sus nodos hay exactamente un camino.

▷ **Grafo múltiplemente conexo.** Es el que contiene ciclos o bucles.

▷ **Árbol.** Es un caso particular de poliárbol, en que cada nodo tiene un sólo padre, excepto el *nodo raíz*, que no tiene padres.

Por ejemplo, el grafo de la figura 3.5 es un poliárbol, porque no contiene bucles; no es un árbol porque algunos de sus nodos (D y H) tienen más de un padre.

3.2.2 Definición de red bayesiana

La propiedad fundamental de una red bayesiana es la *separación direccional* (llamada *d-separation* por Pearl [44, 45]), que se define así:

▷ **Separación direccional.** Dado un grafo dirigido acíclico conexo y una distribución de probabilidad sobre sus variables, se dice que hay *separación direccional* si, dado un nodo X, el conjunto de sus padres, $pa(X)$, separa condicionalmente este nodo de todo otro subconjunto $\bar{Y}$ en que no haya descendientes de X. Es decir,

$$P(x|pa(x), \bar{y}) = P(x|pa(x)) \tag{3.59}$$

Es habitual definir las redes bayesianas a partir de grafos dirigidos acíclicos (en inglés se suelen denominar *directed acyclic graph, DAG*, aunque lo correcto es decir *acyclic directed graph, ADG*). Sin embargo, nos parece importante incluir la especificación "*conexo*" por tres razones. La primera, porque muchos de los algoritmos y propiedades de las redes bayesianas sólo son correctos para grafos conexos, por lo que es mejor incluir esta característica en la definición que tener que añadirla como nota a pie de página en casos particulares. La segunda razón es que, aun en el caso de que tuviéramos un modelo con dos partes inconexas, podríamos tratarlo como dos redes bayesianas independientes. Y la tercera, porque los modelos del mundo real con que vamos a trabajar son siempre conexos; si hubiera dos partes inconexas no tendríamos uno sino dos modelos independientes.

La definición de separación direccional, aunque pueda parecer extraña a primera vista, es sencilla, y ya fue introducida en los ejemplos de la sección 3.1. En efecto, volviendo a la figura 3.4 de dicha sección (pág. 46), recordamos que, una vez conocido el valor de x, podíamos calcular la probabilidad de y_1 sin que influyeran los valores de las demás variables. Es decir, el conjunto $pa(Y_1) = \{X\}$, separa condicionalmente Y_1 de todas las demás variables de la red.

A partir de las definiciones anteriores, podemos caracterizar las redes bayesianas así:

▷ **Red bayesiana.** Es un grafo dirigido acíclico conexo más una distribución de probabilidad sobre sus variables, que cumple la propiedad de separación direccional.

El término *direccional* hace referencia a la asimetría de dicha propiedad, que se manifiesta en las siguientes propiedades de las redes bayesianas, ilustradas con el ejemplo de la figura 3.5:

1. Si A no tiene padres, entonces $P(x|pa(x)) = P(x|\varnothing) = P(x)$, y la ecuación (3.59) se traduce en $P(e|a) = P(e)$ para cada nodo E que no sea uno de los descendientes de A; en otras palabras, E es a priori independiente de A. En consecuencia, dos nodos cualesquiera D y E que no tengan ningún antepasado común son independientes a priori.

2. Si D es descendiente de A y antepasado de H, y no existe ningún otro camino desde A hasta H, entonces estos dos nodos quedan condicionalmente separados por D:

$$P(h|d,a) = P(h|d) \tag{3.60}$$

3. Si tanto G como H son hijos de D y no tienen ningún otro antepasado común, este último separa G y H, haciendo que sean condicionalmente independientes:

$$P(g|d,h) = P(g|d) \tag{3.61}$$

En general, la independencia (a priori o condicional) de dos nodos –por ejemplo, A y E— se pierde al conocer el valor de cualquiera de sus descendientes comunes —H es descendiente tanto de A como de E— pues en este caso la propiedad de separación direccional ya no es aplicable. Es muy importante que observe la relación de estas propiedades con la discusión de la sección 2.2.3.

3.2.3 Factorización de la probabilidad

En la definición de red bayesiana, hemos partido de una distribución de probabilidad conjunta para las variables, $P(\bar{x})$. Aparentemente, en el caso de variables binarias, harían falta $2^N - 1$ parámetros. (Serían 2^N si no existiera la ligadura (2.1).) Sin embargo, las condiciones de independencia dadas por la separación direccional imponen nuevas restricciones, que reducen los grados de libertad del sistema. De hecho, una de las propiedades más importantes de una red bayesiana es que su distribución de probabilidad puede expresarse mediante el producto de las distribuciones condicionadas de cada nodo dados sus padres, tal como nos dice el siguiente teorema. (Recordemos que, para un nodo X sin padres, $pa(X) = \varnothing$ y, por tanto, $P(x|pa(x)) = P(x)$; es decir, la probabilidad condicionada de un nodo sin padres es simplemente la probabilidad a priori.)

Teorema 3.7 (Factorización de la probabilidad) Dada una red bayesiana, su distribución de probabilidad puede expresarse como

$$P(x_1, \ldots, x_n) = \prod_i P(x_i|pa(x_i)) \tag{3.62}$$

Demostración. Es fácil construir una ordenación de las variables en que los padres de cada nodo aparezcan siempre después de él. Supongamos, sin pérdida de generalidad, que la ordenación $\{X_1, \ldots, X_n\}$ cumple dicha propiedad. Por la proposición 2.10 (ec. (2.14)), podemos escribir

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|x_{i+1}, \ldots, x_n)$$

Ahora bien, por la forma en que hemos escogido la ordenación, el conjunto $\{X_{i+1}, \ldots, X_n\}$ incluye todos los padres de X y, en consecuencia, la separación direccional (ec. (3.59)) nos dice que

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

con lo que concluimos la demostración. □

Ejemplo 3.8 Para la red bayesiana de la figura 3.4 (pág. 46), la factorización de la probabilidad viene dada por la ecuación (3.58).

Ejemplo 3.9 Para el grafo de la figura 3.5 (pág. 49), la factorización de la probabilidad viene dada por

$$\begin{aligned} P(a,b,c,d,e,f,g,h,i) &= P(a)\cdot P(b)\cdot P(c)\cdot P(d|a,b)\cdot P(e|c) \\ &\quad \cdot P(f|c)\cdot P(g|d)\cdot P(h|d,e)\cdot P(i|e) \end{aligned}$$

Podemos comprobar que cada uno de estos factores corresponde a una de las familias enumeradas en el ejemplo 3.5.

La importancia de este teorema es que nos permite describir una red bayesiana a partir de la probabilidad condicionada de cada nodo, en vez de dar la distribución de probabilidad conjunta, que requeriría un número de parámetros exponencial en el número de nodos y plantearía el grave problema de verificar la propiedad de separación direccional; sin embargo, el número de parámetros requerido para dar las probabilidades condicionadas es proporcional al número de nodos (suponiendo que el número de padres y el número de valores posibles están acotados para cada variable).

Podríamos haber definido las propiedades de independencia en términos de caminos activados o bloqueados, al estilo de Pearl, Geiger y Verma [45, págs. 317-318], [47], seguido también por Charniak [7]. En cambio, la presentación que hemos escogido se parece más a la propuesta por Neapolitan [41, cap. 5].

3.2.4 Semántica de las redes bayesianas

Hemos definido ya las redes bayesianas desde un punto de vista matemático formal. La cuestión que nos planteamos ahora es su semántica, es decir, ¿qué interpretación se le puede dar a una red bayesiana? ¿Cómo se corresponde nuestro modelo con el mundo real? ¿Por qué podemos hablar de causas y efectos en una R.B.?

Esta cuestión está ya parcialmente respondida en la sección 3.1, que fue introducida antes de la definición formal de R.B. precisamente para mostrar que los conceptos y axiomas introducidos no pareciesen arbitrarios, sino que responden a las propiedades de la causalidad, según nuestra concepción intuitiva del mundo real.

Es importante señalar que la estructura de la red, por sí misma, aporta gran cantidad de información cualitativa. En efecto, un arco XY indica, ya antes de conocer el valor concreto de la probabilidad condicional, que hay una correlación entre ambas variables: el valor que toma X influye sobre la probabilidad de Y, y viceversa. Es lo que llamamos *influencia causal directa*. Tal es la relación que existe, por ejemplo, entre el país de origen y el paludismo, o entre el paludismo y la fiebre. Profundizando un poco más, observamos que la existencia de un camino entre dos variables X e Y, con variables intermedias $\bar{Z}$, indica que hay una *influencia causal indirecta* entre ambas.

Tal como hemos discutido en la presentación intuitiva de las RR.BB., cuando nuestro sentido común, basado en la experiencia, nos dice que la influencia de una variable X sobre uno de sus efectos Y_1 (por ejemplo, del paludismo sobre la prueba de la gota gruesa) no depende

de cuáles han sido las causas o mecanismos que han producido X, ni depende tampoco de si X a dado lugar a otros efectos, entonces la red contendrá un arco desde X hasta Y_1, y no habrá ningún arco que conecte Y_1 con las demás variables. Por tanto, *la ausencia de arcos* es también una forma de expresar información. El hecho de que Y_1 depende solamente de su causa, X, se traduce matemáticamente diciendo que, conocido el valor de X, la probabilidad de Y_1 es independiente de los valores que toman esas otras variables, o dicho de otro modo, X separa Y_1 de dichas variables. Empezamos a ver aquí la relación entre el concepto de padre y el de causa, entre el de hijo y el de efecto, entre el de arco y el de influencia causal directa, entre el de independencia en los mecanismos causales y el de independencia probabilista.

En este punto es donde se manifiesta la importancia del sentido de los arcos y su relación con la idea de causalidad. Volviendo al ejemplo del paludismo, el hecho de que las variables País-de-origen y Tipo-sanguíneo no tengan ningún padre común significa que son a priori independientes, es decir, que el país no influye en el tipo sanguíneo y viceversa, de modo que, si no hay más evidencia, no podemos obtener ninguna información sobre el país de origen a partir del tipo sanguíneo, ni viceversa. Sin embargo, el hecho de que ambas variables tengan un hijo común significa que, una vez conocido el valor de ese nodo, surgen correlaciones entre los padres.[7] Podemos decir, usando la terminología de Pearl [44], que el camino entre U_1 y U_2 permanece *bloqueado* hasta que sea *activado* por la llegada de información sobre X o sobre alguno de sus descendientes.

Para el caso de los efectos de una variable ocurre precisamente lo contrario: todo médico sabe que hay correlación entre la fiebre y el test de la gota gruesa. Sin embargo, tal como discutimos en la sección 3.1, la correlación desaparece cuando averiguamos si el paciente tiene o no tiene paludismo. Es decir, el camino entre Y_1 e Y_2 está *activado* en principio, y *se bloquea* sólo al conocer el valor de X. De esta asimetría entre padres e hijos, reflejo de la asimetría que existe en el mundo real entre causas y efectos, procede el nombre de separación *direccional*.

Por tanto, hay dos formas de justificar los enlaces que introducimos u omitimos al construir nuestra red. La primera es de naturaleza teórica: formamos un modelo causal a partir de la experiencia de un especialista y trazamos los arcos correspondientes al modelo; la relación que hemos discutido entre los mecanismos causales y las propiedades matemáticas de independencia nos permite fundamentar nuestro modelo. El otro camino para justificar la red consiste en realizar una comprobación empírica a partir de un conjunto suficientemente amplio de casos, utilizando las herramientas estadísticas que se emplean habitualmente para detectar correlaciones.

Hay otro punto relativo a la semántica de las redes bayesianas, que vamos a mencionar sólo brevemente, pues aún está muy discutido. Nos referimos al debate entre los que defienden que las redes probabilistas pueden expresar causalidad y los que sostienen que éstas sólo expresan correlaciones entre variables. En realidad, no se trata de un debate limitado al campo de las redes bayesianas, sino que la existencia de la causalidad es una cuestión que se han planteado matemáticos y filósofos por lo menos desde el siglo XVII, a partir de las teorías de Hume. Para no entrar en esta cuestión citaremos solamente tres trabajos, los de Pearl y Verma [48, 46] y el de Druzdzel y Simon [19], que muestran cómo recientemente han surgido argumentos matemáticos para defender la interpretación causal frente a la meramente correlacional.

En resumen, lo que hemos intentado mostrar en esta sección es que la información cualitativa que expresa la estructura de una R.B. es más importante aún que la información

[7] La correlación que aparece entre las causas se aprecia mucho más claramente en el caso de la puerta OR (sec. 3.4).

cuantitativa, como lo demuestra el hecho de que se han construido *redes cualitativas* [64, 65], capaces de razonar a partir de las propiedades de independencia de las redes bayesianas, incluso en ausencia de valores numéricos. Por este motivo, Neapolitan [41] ha sugerido en nombre de *redes de independencia* (*independence networks*) como el más adecuado para las RR.BB.

Podríamos sintetizar todo lo dicho anteriormente repitiendo lo que Laplace afirmó en la introducción de su famoso libro *Théorie Analytique des Probabilités*:[8]

> La teoría de la probabilidad no es, en el fondo, más que el sentido común reducido al cálculo.

3.3 Propagación de evidencia en poliárboles

Vamos a estudiar ahora un algoritmo eficiente para calcular la probabilidad en una red bayesiana sin bucles. En realidad, dada una R.B., a partir de las probabilidades condicionales podríamos calcular la probabilidad conjunta según el teorema 3.7, y luego aplicar las ecuaciones (2.2) y (2.6) para calcular las probabilidades marginales y a posteriori, respectivamente. Sin embargo este método tendría complejidad exponencial incluso en el caso de poliárboles. Además, al añadir nueva evidencia tendríamos que repetir casi todos los cálculos. Por esta razón conviene encontrar algoritmos mucho más eficientes.

El algoritmo para poliárboles que presentamos en esta sección, basado en el paso de mensajes π y λ, fue desarrollado por Kim [34] a partir del que Pearl había propuesto para árboles [43]. Sin embargo, la principal limitación del algoritmo de Kim y Pearl es que no permite tratar los bucles que aparecen inevitablemente al desarrollar modelos del mundo real, por lo que en sí mismo resulta de muy poca utilidad y los constructores de RR.BB. recurren a otros que, aun perdiendo las ventajas de éste, son aplicables a todo tipo de RR.BB. Sin embargo, aquí lo vamos a estudiar con detalle por dos razones. Primera, por su sencillez y elegancia, que nos permitirán comprender mejor las propiedades de las RR.BB. Y segunda, porque el algoritmo de condicionamiento local [15], aplicable también a redes múltiplemente conexas, es una extensión de éste, que se basa en los mismos conceptos y definiciones.

Para comprender mejor el desarrollo matemático que vamos a realizar, puede ser útil al lector repasar la sección 3.1, en que aparecen sencillos ejemplos numéricos que explican por qué se introducen las definiciones de π y λ, y cómo se propaga la evidencia.

3.3.1 Definiciones básicas

Una de las propiedades fundamentales de un poliárbol es que hay un único camino entre cada par de nodos. En consecuencia, la influencia de cada hallazgo se propaga hasta un nodo X bien a través de los padres o a través de los hijos de éste, por lo que para cada nodo X podemos hacer una partición de la evidencia (recordamos que la evidencia es el conjunto de hallazgos) en dos subconjuntos, tales que

$$\mathbf{e} = \mathbf{e}_X^+ \cup \mathbf{e}_X^- \tag{3.63}$$

$$\mathbf{e}_X^+ \cap \mathbf{e}_X^- = \varnothing \tag{3.64}$$

[8] Cita tomada de Druzdzel [18].

donde $\mathbf{e}_X^+$ representa la evidencia "por encima de X" y $\mathbf{e}_X^-$ "por debajo de X" en el sentido antes mencionado.

De forma similar, la eliminación de un enlace XY divide a la red —y por tanto también la evidencia— en dos partes, una que queda "por encima" del enlace y otra que queda "por debajo". Las llamaremos $\mathbf{e}_{XY}^+$ y $\mathbf{e}_{XY}^-$, respectivamente. Al igual que en el caso anterior, se cumple que

$$\mathbf{e} = \mathbf{e}_{XY}^+ \cup \mathbf{e}_{XY}^- \tag{3.65}$$

$$\mathbf{e}_{XY}^+ \cap \mathbf{e}_{XY}^- = \varnothing \tag{3.66}$$

Ejemplo 3.10 En la figura 3.5 (pág. 49), si tuviéramos $\mathbf{e} = \{+f, +g, \neg i\}$, entonces $\mathbf{e}_E^+ = \{+f\}$ y $\mathbf{e}_E^- = \{+g, \neg i\}$. Del mismo modo, $\mathbf{e}_H^+ = \{+f, +g, \neg i\}$ y $\mathbf{e}_H^- = \varnothing$. La eliminación del enlace EH dividiría la red en dos partes, y tendríamos $\mathbf{e}_{EH}^+ = \{+f, \neg i\}$ y $\mathbf{e}_{EH}^- = \{+g\}$. □

Basándonos en la partición de la evidencia, podemos establecer las siguientes definiciones (cf. fig. 3.7):

$$\pi(x) \equiv P(x, \mathbf{e}_X^+) \tag{3.67}$$

$$\lambda(x) \equiv P(\mathbf{e}_X^- \,|\, x) \tag{3.68}$$

$$\pi_X(u_i) \equiv P(u_i, \mathbf{e}_{U_i X}^+) \tag{3.69}$$

$$\lambda_{Y_j}(x) \equiv P(\mathbf{e}_{XY_j}^- \,|\, x) \tag{3.70}$$

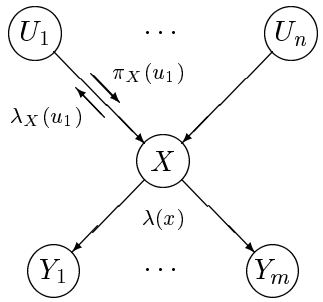

Figura 3.7: Propagación de evidencia mediante intercambio de mensajes.

El sentido de estas definiciones es el siguiente:

- $\pi(x)$ indica qué valor de X es más probable según la evidencia relacionada con las causas de X (es decir, según la evidencia "por encima" de X).
- $\lambda(x)$ indica qué valor de X explica mejor los hallazgos correspondientes a los efectos de X (la evidencia "por debajo" de X).
- $\pi_X(u)$ indica qué valor de U es más probable según la evidencia "por encima" del enlace UX.
- $\lambda_{Y_j}(x)$ indica qué valor X explica mejor la evidencia "por debajo" del enlace XY.

Para entender mejor estas explicaciones, conviene volver a los ejemplos de la sección 3.1.

Antes de concluir esta sección, señalemos que las definiciones anteriores, aunque tomadas del libro de Pearl [45], han sido modificadas de acuerdo con la propuesta de Peot y Shachter [50], con el fin de permitir un tratamiento coherente de los bucles mediante el algoritmo de condicionamiento local [15].

3.3.2 Computación de los mensajes

Recordemos una vez más que nuestro objetivo es calcular la probabilidad a posteriori de cada nodo, definida en la ecuación (2.25). A partir, de ahí,

$$\begin{aligned} P^*(x) = P(x|\,\mathbf{e}) &= \alpha\, P(x, \mathbf{e}_X^+, \mathbf{e}_X^-) \\ &= \alpha\, P(x, \mathbf{e}_X^+)\, P(\mathbf{e}_X^-|\, x, \mathbf{e}_X^+) \end{aligned}$$

donde hemos definido

$$\alpha \equiv [P(\mathbf{e})]^{-1} \tag{3.71}$$

Ahora bien, por la separación direccional sabemos que $P(\mathbf{e}_X^-|\, x, \mathbf{e}_X^+) = P(\mathbf{e}_X^-|\, x)$, de modo que, aplicando las definiciones anteriores llegamos a

$$P^*(x) = \alpha\, \pi(x)\, \lambda(x) \tag{3.72}$$

Necesitamos, por tanto, calcular los tres factores que aparecen en esta expresión. Empecemos con $\pi(x)$. Según su definición,

$$\pi(x) = P(x, \mathbf{e}_X^+) = \sum_{\bar{u}} P(x|\,\bar{u}) P(\bar{u}, \mathbf{e}_X^+)$$

Como las causas de X no tienen ningún antepasado común por estar en un poliárbol (red simplemente conexa), todas ellas y las ramas correspondientes son independientes mientras no consideremos la evidencia relativa a X o a sus descendientes:

$$\begin{aligned} P(\bar{u}, \mathbf{e}_X^+) &= P(u_1, \mathbf{e}_{U_1X}^+, \ \ldots\ , u_n, \ \mathbf{e}_{U_nX}^+) \\ &= \prod_{i=1}^{n} P(u_i, \mathbf{e}_{U_iX}^+) = \prod_{i=1}^{n} \pi_X(u_i) \end{aligned} \tag{3.73}$$

Por tanto,

$$\pi(x) = \sum_{\bar{u}} P(x|\,\bar{u}) \prod_{i=1}^{n} \pi_X(u_i)\,. \tag{3.74}$$

El paso siguiente consiste en calcular $\pi_X(u_i)$ o, lo que es lo mismo, $\pi_{Y_j}(x)$, puesto que en una R.B. todos los nodos son equivalentes; es sólo una cuestión de notación. La evidencia que está por encima del enlace XY_j, $\mathbf{e}_{XY_j}^+$, podemos descomponerla en varios subconjuntos: la que está por encima de X y la que está por debajo de cada enlace XY_k para los demás efectos Y_k de X (fig. 3.7). Sabemos además que X separa $\mathbf{e}_X^+$ de $\mathbf{e}_{XY_k}^-$, y separa también los

subconjuntos $\mathbf{e}^-_{XY_k}$ entre sí. Con estas consideraciones, obtenemos

$$\begin{aligned}\pi_{Y_j}(x) &= P(x, \mathbf{e}^+_{XY_j}) = P(x, \mathbf{e}^+_X, \mathbf{e}^-_{XY_k}\ k\neq j)\\ &= P(x, \mathbf{e}^+_X) \prod_{k\neq j} P(\mathbf{e}^-_{XY_k}|\, x)\\ &= \pi(x) \prod_{k\neq j} \lambda_{Y_k}(x)\end{aligned} \tag{3.75}$$

Para calcular esta expresión, es necesario hallar $\lambda_{Y_k}(x)$ —o $\lambda_{Y_j}(x)$, pues el resultado obtenido será válido para todos los efectos de X—. Representaremos mediante $\bar{V}$ el conjunto de causas de Y_j (o del efecto considerado) distintas de X, tal como muestra la figura 3.8. Por simplificar la notación, escribiremos $\mathbf{e}^+_{\bar{V}Y_j} = \mathbf{e}^+_{V_1Y} \cup \ldots \cup \mathbf{e}^+_{V_pY}$, con lo que nos queda $\mathbf{e}^-_{XY_j} = \mathbf{e}^-_Y \cup \mathbf{e}^+_{\bar{V}Y}$.

Recordemos que Y_j separa $\mathbf{e}^-_{Y_j}$ del resto de la red que está por encima de Y_j, e igualmente los padres de Y_j separan Y_j de $\mathbf{e}^+_{\bar{V}Y_j}$. Aplicando repetidamente la proposición 2.19, resulta

$$\begin{aligned}\lambda_{Y_j}(x) &= P(\mathbf{e}^-_{XY_j}|\, x)\\ &= \sum_{y_j}\sum_{\bar{v}} P(\mathbf{e}^-_{Y_j}, y_j, \mathbf{e}^+_{\bar{V}Y_j}, \bar{v}|\, x)\\ &= \sum_{y_j}\sum_{\bar{v}} P(\mathbf{e}^-_{Y_j}|\, y_j)\, P(y_j|\, \bar{v}, x)\, P(\mathbf{e}^+_{\bar{V}Y_j}, \bar{v}|\, x)\end{aligned}$$

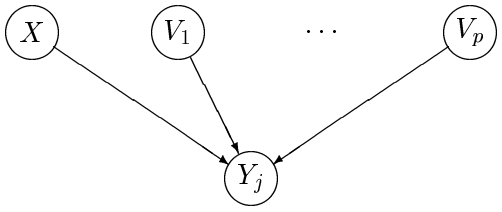

Figura 3.8: Padres de Y_j.

Puesto que las causas de Y_j son independientes a priori, podemos razonar como en la ecuación (3.73) para llegar a

$$P(\bar{v}, \mathbf{e}^+_{\bar{V}Y_j}|\, x) = P(\bar{v}, \mathbf{e}^+_{\bar{V}Y_j}) = \prod_{l=1}^{p} P(v_l, \mathbf{e}^+_{V_lY_j}) = \prod_{l=1}^{p} \pi_{Y_j}(v_l)$$

y, en consecuencia,

$$\lambda_{Y_j}(x) = \sum_{y_j} \left[\lambda(y_j) \sum_{\bar{v}} P(y_j|\, x, \bar{v}) \prod_{l=1}^{p} \pi_{Y_j}(v_l)\right] \tag{3.76}$$

Finalmente, hay que calcular $\lambda(x)$, lo cual resulta bastante sencillo:

$$\begin{aligned}\lambda(x) &= P(\mathbf{e}^-_{XY_1}, \ldots, \mathbf{e}^-_{XY_m}|\, x)\\ &= \prod_{j=1}^{m} P(\mathbf{e}^-_{XY_j}|\, x) = \prod_{j=1}^{m} \lambda_{Y_j}(x)\end{aligned} \tag{3.77}$$

Para completar el algoritmo, falta hallar la constante α que aparece en (3.72). Realizar el cálculo a partir de la definición 3.71 resultaría muy complicado en general. Sin embargo, sabemos que

$$\sum_x P^*(x) = \alpha \sum_x \pi(x)\ \lambda(x) = 1 \tag{3.78}$$

con lo que podemos obtener α como

$$\alpha = \left[\sum_x \pi(x)\ \lambda(x) \right]^{-1} \tag{3.79}$$

En la práctica, calcularemos $\pi(x)$ y $\lambda(x)$ para cada nodo y normalizaremos su producto de acuerdo con la ecuación (3.78).

Observe que por cada enlace $X \to Y$ circulan dos mensajes, $\pi_Y(x)$ de X a Y, y $\lambda_Y(x)$, de Y a X, pero ambos mensajes son vectores correspondientes a la variable X (por tanto, la dimensión del vector es $|X|$, el número de valores de X), mientras que la variable Y sólo aparece como subíndice en los dos: en $\pi_Y(x)$ indica el nodo que recibe el mensaje, mientras que en $\lambda_Y(x)$ indica el que lo envía.

3.3.3 Comentarios

Las fórmulas que acabamos de deducir son recursivas: $\pi(x)$ se calcula a partir de $\pi_X(u_i)$; $\pi_{Y_j}(x)$ a partir de $\pi(x)$ y de $\lambda_{Y_k}(x)$, etc. Necesitamos por tanto una condición de terminación para que el algoritmo esté completo. Por otro lado, necesitamos explicar cómo introducir en este esquema la evidencia observada. Resolveremos ambos problemas del siguiente modo:

Para un nodo U sin padres, $\mathbf{e}^+_U = \varnothing$, por lo que $\pi(u) = P(u)$, que es uno de los parámetros que definen la red. En este caso el problema de terminación ya lo teníamos resuelto.

Para un nodo terminal Y (nodo sin hijos), hace falta conocer $\lambda(y)$. Si no hay ninguna información sobre este nodo, asignamos el mismo número para cada valor y; por ejemplo, $\lambda(y) = 1$ para todo y. Vemos en la ecuación (3.72) que un vector $\lambda(x)$ constante no modifica el valor de $P^*(x)$. También vemos, a partir de la ecuación (3.76), que para un vector constante $\lambda(y) = 1$ podemos alterar el orden de los sumatorios y llegar a

$$\begin{aligned} \lambda_{Y_j}(x) &= c \sum_{\bar{v}} \left[\prod_{l=1}^{p} \pi_{Y_j}(v_l) \sum_{y_j} P(y_j \,|\, x, \bar{v}) \right] \\ &= c \sum_{\bar{v}} \left[\prod_{l=1}^{p} \pi_{Y_j}(v_l) \right] = c \sum_{\bar{v}} P(\bar{v}, \mathbf{e}^+_{\bar{V}Y_j}) = c\, P(\mathbf{e}^+_{\bar{V}Y_j}), \quad \forall x \end{aligned} \tag{3.80}$$

que es de nuevo un vector constante —es decir, independiente de x— y no transmite ninguna información, pues según las ecuaciones (3.72) y (3.76), un vector λ constante no influye en el resultado final.

Si hay un nodo terminal Y de valor conocido y_0 (es decir, la afirmación "$Y = y_0$" es parte de la evidencia), asignamos a $\lambda(y_0)$ un número positivo cualquiera y 0 a los demás valores de Y. Por ejemplo,

$$\begin{cases} \lambda(y_0) = 1 \\ \lambda(y) \ = 0 \quad \text{para } y \neq y_0 \end{cases}$$

lo cual implica, según (3.72),

$$\begin{cases} P(y_0) = 1 \\ P(y) \ = 0 \quad \text{para } y \neq y_0 \end{cases}$$

Vemos que, efectivamente, la probabilidad se ajusta a la afirmación de partida, "$Y = y_0$"; además sólo el valor y_0 cuenta en el sumatorio de la ecuación (3.76), por lo que podemos concluir que esta asignación de $\lambda(y)$ para nodos terminales es coherente, y así queda completo el algoritmo de propagación de evidencia en poliárboles.

En resumen, para que un algoritmo recursivo sea correcto debe haber una condición de terminación, que en este caso se satisface por la inicialización del algoritmo, al asignar una π a cada nodo sin padres y una λ a cada nodo sin hijos. Esto explica también por qué este algoritmo sólo se puede aplicar a poliárboles, pues si la red tuviera bucles este algoritmo solicitaría una y otra vez los mismos mensajes, y no terminaría nunca.[9]

Ejemplo 3.11 Volvamos de nuevo a la red de la figura 3.5 (pág. 49). Recordemos que, además de tener la estructura de la red, conocemos las probabilidades a priori de los nodos sin padres: $P(a)$, $P(b)$ y $P(c)$, y las probabilidades condicionales: $P(d|\, a, b)$, $P(e|\, c)$, etc.

Supongamos que $\mathbf{e} = \{+f, +g, \neg i\}$. La asignación de landas para los nodos terminales será:

$$\begin{cases} \lambda(+f) = 1 \\ \lambda(\neg f) \ = 0 \end{cases} \quad \begin{cases} \lambda(+g) \ = 1 \\ \lambda(\neg g) \ = 0 \end{cases} \quad \begin{cases} \lambda(+h) = 1 \\ \lambda(\neg h) \ = 1 \end{cases} \quad \begin{cases} \lambda(+i) \ = 0 \\ \lambda(\neg i) \ = 1 \end{cases}$$

Queremos calcular $P^*(e)$, y por eso escogemos el nodo E como *pivote*, en el sentido de que se va a encargar de solicitar información a todos sus vecinos. Es posible que luego otros nodos soliciten los mensajes que les faltan, con el fin de computar su propia probabilidad, aunque también es posible que el nodo pivote E, una vez que ha recibido todos sus mensajes "decida" computar y enviar los mensajes de vuelta para sus vecinos, con el fin de que éstos hagan lo mismo con sus demás vecinos, y así sucesivamente hasta alcanzar todos los nodos terminales del poliárbol.

Con este esquema en mente, empezamos buscando $\pi(e)$:

$$\begin{aligned} \pi(e) &= \sum_c P(e|\, c)\, \pi_E(c) \\ \pi_E(c) &= \pi(c)\, \lambda_F(c) = P(c)\, \lambda_F(c) \\ \lambda_F(c) &= \sum_f \lambda(f)\, P(f|\, c) = P(+f|\, c) \end{aligned}$$

Así concluimos el cálculo en esta rama del árbol. Continuamos con otras ramas:

$$\begin{aligned} \lambda(e) &= \lambda_I(e)\, \lambda_H(e) \\ \lambda_I(e) &= \sum_i \lambda(i)\, P(i|\, e) \ = P(\neg i|\, e) \\ \lambda_H(e) &= \sum_h \lambda(h) \sum_d P(h|\, d, e)\, \pi_H(d) \end{aligned}$$

[9] En realidad, la verdadera razón por la que este algoritmo sólo sirve para poliárboles es que se fundamenta en la hipótesis de que la red no tiene bucles, y por eso el algoritmo no se podría aplicar a redes con bucles aunque consiguiéramos satisfacer de alguna manera la condición de terminación.

Deberíamos calcular ahora $\pi_H(d)$. Sin embargo, podemos saber ya que $\lambda_H(e)$ va a ser un vector constante porque $\lambda(h)$ también lo es. Podemos demostrarlo mediante el argumento numérico de la ecuación (3.80). Otra forma de razonarlo es a partir de las propiedades de independencia condicional: cuando el valor de H no se conoce, D y E son independientes; recordando además que D separa G de E, tenemos

$$\begin{aligned}\lambda_H(e) &= P(\mathbf{e}^-_{EH} \,|\, e) = P(+g \,|\, e) \\ &= \sum_d P(+g \,|\, d, e)\, P(d \,|\, e) = \sum_d P(+g \,|\, d)\, P(d) = P(+g)\end{aligned}$$

que es un vector constante (no depende de e).

Por fin, nos queda $\lambda(e) = \lambda_I(e)$ y basta normalizar el producto $\pi(e) \cdot \lambda(e)$ para conocer $P^*(x)$. Del mismo modo podemos calcular la probabilidad a posteriori de cualquier otra variable, aprovechando —naturalmente— los resultados ya obtenidos. □

3.3.4 Implementación distribuida

El algoritmo que hemos presentado se presta inmediatamente a una implementación recursiva, según hemos comentado anteriormente. Vamos a ver ahora cómo podemos diseñar un algoritmo distribuido a partir de las mismas expresiones. (Veremos también que este método puede llevarnos a una implementación iterativa, que presenta la ventaja de que requiere mucha menos memoria de cálculo que la implementación recursiva.)

En la implementación distribuida, cada procesador corresponderá a un nodo y, por tanto, a una variable. La información que debe almacenar puede ser estática o dinámica. Aquí, "estática" significa "independiente de la evidencia observada", tal como la estructura de la red y las probabilidades condicionales. En caso de que los nodos no sean procesadores físicos reales sino que estén simulados mediante un programa de ordenador, lo primero que cada nodo necesita conocer son sus causas y sus efectos; esto puede realizarse fácilmente definiendo dos listas con punteros hacia los nodos correspondientes, las cuales codifican la topología de la red. A continuación hay que introducir la información numérica estática, a saber, las probabilidades a priori y condicionales.

La información dinámica consiste, en primer lugar, en los valores de λ para los nodos terminales, tal como hemos explicado anteriormente, y en los mensajes π y λ correspondientes a la propagación de evidencia. La propiedad más importante que se deriva de los axiomas de independencia descritos en el capítulo anterior es que, en poliárboles, cada enlace descompone la red en dos partes cuya única interacción se transmite a través de dicho enlace, y los mensajes intercambiados están desacoplados, en el sentido de que $\pi_Y(x)$ puede calcularse independientemente de $\lambda_Y(x)$, y viceversa. En la figura 3.9, que muestra los cálculos realizados en el nodo X, esta propiedad aparece como la ausencia de bucles en el flujo de información. Se puede comprobar también, observando las fórmulas de la sección 3.3.2, que toda la información requerida por un nodo para computar sus mensajes se encuentra almacenada localmente.

Un nodo X está en disposición de enviar un mensaje a su vecino W cuando y sólo cuando ha recibido ya los mensajes procedentes de todos sus demás vecinos. Un nodo X con n causas y m efectos que ha recibido q mensajes se encuentra en uno de tres estados posibles:

1. $q \leq n + m - 2$. Esto significa que X está esperando al menos dos mensajes, por lo que todavía no puede calcular ninguno de los que debe enviar.

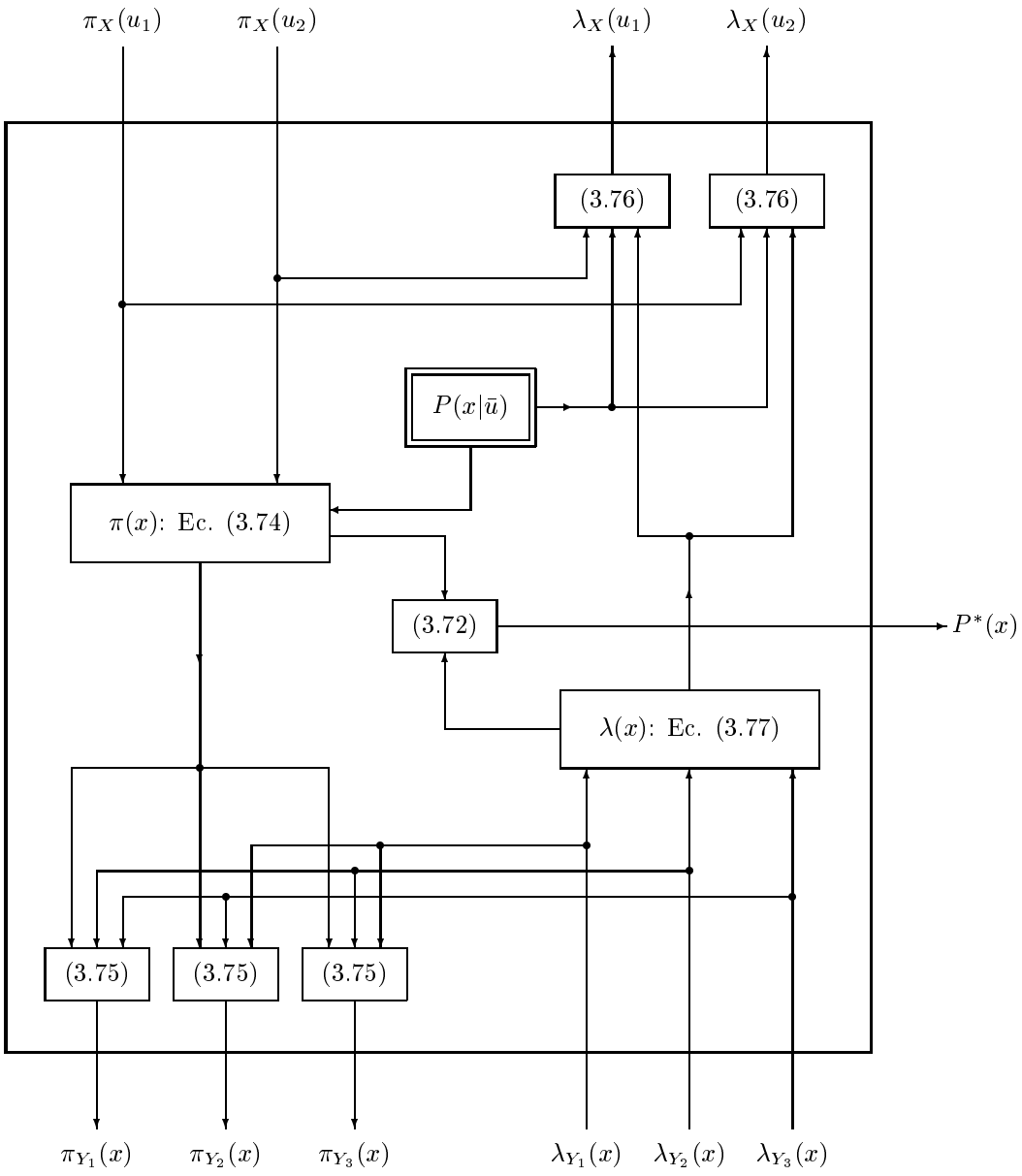

Figura 3.9: Computaciones realizadas en el nodo X.

2. $q = n + m - 1$. En este caso, X ha recibido un mensaje de cada vecino excepto de uno, que llamaremos W. Por eso X puede calcular ya el mensaje que debe enviar a W (aunque todavía no puede calcular ningún otro mensaje).

3. $q = n + m$. Cuando X ha recibido todos los mensajes que estaba esperando, puede calcular por fin los que le faltaban por enviar.

Al principio, $q = 0$ para todos los nodos, pues aún no ha circulado ningún mensaje; por tanto, todos los nodos con un solo vecino ($n+m=1$) se encuentran en el estado 2; los demás se encuentran todavía en el estado 1. Es posible demostrar que siempre hay algún nodo dispuesto a enviar un mensaje, por lo que el proceso no se interrumpe nunca hasta que el algoritmo se ha completado. En vez de realizar la demostración, que es sencilla conceptualmente pero engorrosa, volvamos una vez más a la figura 3.5.

Antes de que empiece la propagación, todos los nodos que tienen un solo vecino (A, B, F, G y I) se hallan en estado 2, y los demás en estado 1. Cuando aquéllos envían sus mensajes respectivos, C y D pasan al estado 2, y lo mismo ocurre en el paso siguiente con E y H. Cuando los mensajes $\pi_H(e)$ y $\lambda_H(e)$ llegan a su destino, estos dos últimos nodos pasan al estado 3, de modo que pueden enviar ya a sus vecinos los mensajes que faltaban, y en dos pasos más queda concluido el proceso.

La discusión anterior es interesante para demostrar que no es necesario tener un mecanismo global de control, por lo que el modelo puede implementarse como una *red asíncrona* en que el número de mensajes recibido determina qué mensajes puede calcular y enviar cada nodo.

Si el algoritmo se implementa secuencialmente y la computación necesaria en cada nodo está acotada (limitando el número de padres y valores), el tiempo de computación es proporcional al número de nodos. En este caso resulta más eficiente realizar la propagación de evidencia en dos fases: recolección de mensajes hacia el nodo pivote y distribución desde él, como propusieron Jensen, Olesen y Andersen [32] para árboles de cliques.

En cambio, si hay un procesador por cada nodo, el tiempo de computación es proporcional a la longitud máxima que exista dentro de la red. La versión que hemos presentado aquí, basada en tres estados diferentes para cada nodo, se diferencia ligeramente de la de Pearl [45] en que evita computar y enviar mensajes prematuros carentes de sentido. La distinción no tiene importancia si disponemos de un procesador físico (*hardware*) por cada nodo. Pero si los procesadores conceptuales (los nodos) están simulados por un número menor de procesadores reales, el despilfarro computacional de enviar mensajes inútiles puede resultar muy caro en términos de eficiencia. En este último caso, en que los nodos hacen cola para acceder a un número limitado de procesadores físicos, encontramos el problema típico de la programación distribuida, a saber, cuál de los mensajes debe computarse primero con el fin de lograr la máxima eficiencia.

Ejemplo 3.12 Sea una red bayesiana dada por el grafo de la figura 3.10 y por las siguientes tablas de probabilidad (suponemos que todas las variables son binarias, de modo que $P(\neg a) = 1 - P(+a)$, etc.):

$$P(+a) = 0'3 \qquad P(+b) = 0'1$$

$$\left\{ \begin{array}{cc} P(+c|+a,+b) = 0'9 & P(+c|+a,\neg b) = 0'2 \\ P(+c|\neg a,+b) = 0'3 & P(+c|\neg a,\neg b) = 0'1 \end{array} \right\}$$

$$\{P(+d|+b) = 0'8 \qquad P(+d|\neg b) = 0\}$$

Dada la evidencia $\mathbf{e} = \{+a, \neg d\}$, calcular todos los mensajes π y λ que intervienen y la probabilidad a posteriori de cada variable.

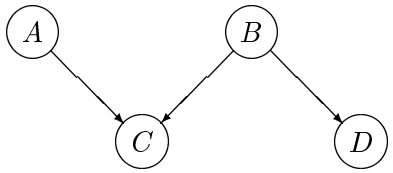

Figura 3.10: Computación distribuida de los mensajes π y λ.

Solución. Empezamos por asignar la evidencia observada. Así, el hallazgo $+a$ implica que $\pi(a) = (1\ 0)$ —este vector significa que $\pi(+a) = 1$ y $\pi(\neg a) = 0$— mientras que el hallazgo $\neg d$ se traduce en $\lambda(d) = (0\ 1)$. Como B no tiene evidencia ni directa ni procedente de sus padres, $\pi(b) = P(b) = (0'1\ 0'9)$; es decir, le asignamos su probabilidad a priori. Al nodo C le asignamos un vector constante, $\lambda(c) = (1\ 1)$, porque no tiene evidencia asociada directamente ni procedente de sus hijos.

Ahora hay que empezar a propagar la evidencia, de acuerdo con las ecuaciones (3.74) a (3.77). Vemos que A está esperando un solo mensaje, $\lambda_C(a)$, y D está esperando un solo mensaje, $\pi_D(b)$, de modo que están ya en condiciones de empezar a enviar mensajes. En cambio B y C están esperando dos mensajes cada uno, por lo que todavía no pueden enviar ninguno.

El mensaje que envía A es $\pi_C(a) = \pi(a) = (1\ 0)$, porque A no tiene otros hijos. El que envía D es $\lambda_D(b) = \sum_d \lambda(d) \cdot P(d|b) = 0 \cdot P(d|b) + 1 \cdot P(\neg d|b) = P(\neg d|b) = 1 - P(+d|b) = (0'2\ 1)$. Ahora tanto a B como a C sólo les falta recibir un mensaje, por lo que ya pueden empezar a enviar algunos mensajes.

El nodo B envía el mensaje $\pi_C(b) = \pi(b) \cdot \lambda_D(b) = (0'1\ 0'9) \cdot (0'2\ 1) = (0'02\ 0'9)$, mientras que el nodo C envía el mensaje $\lambda_C(b) = \sum_c [\lambda(c) \cdot \sum_a P(c|a,b) \cdot \pi_C(a)]$. Este mensaje se calcula así:

$$\begin{aligned}
\lambda_C(+b) &= \lambda(+c) \cdot [P(+c|+a,+b) \cdot \pi_C(+a) + P(+c|\neg a,+b) \cdot \pi_C(\neg a)] \\
&\quad + \lambda(\neg c) \cdot [P(\neg c|+a,+b) \cdot \pi_C(+a) + P(\neg c|\neg a,+b) \cdot \pi_C(\neg a)] \\
&= [0'9 \cdot 1 + 0'3 \cdot 0] + [0'1 \cdot 1 + 0'7 \cdot 0] = 1 \\
\lambda_C(\neg b) &= \lambda(+c) \cdot [P(+c|+a,\neg b) \cdot \pi_C(+a) + P(+c|\neg a,\neg b) \cdot \pi_C(\neg a)] \\
&\quad + \lambda(\neg c) \cdot [P(\neg c|+a,\neg b) \cdot \pi_C(+a) + P(\neg c|\neg a,\neg b) \cdot \pi_C(\neg a)] \\
&= [0'1 \cdot 1 + 0'7 \cdot 0] + [0'9 \cdot 1 + 0'3 \cdot 0] = 1
\end{aligned}$$

Por tanto, $\lambda_C(b) = (1\ 1)$; es decir, se trata de un vector constante, que no aporta información. En realidad, el hecho de que $\lambda(c)$ es un vector constante nos permite calcular el mensaje $\lambda_C(b)$ de forma más sencilla que como acabamos de hacerlo:

$$\lambda_C(b) = \sum_c \left[1 \cdot \sum_a P(c|a,b) \cdot \pi_C(a) \right] = \sum_a \left[\sum_c P(c|a,b) \right] \cdot \pi_C(a) = \sum_a \pi_C(a)$$

Esto explica por qué $\lambda_C(+b) = \lambda_C(\neg b)$, es decir, el mensaje $\lambda_C(b)$ es un vector constante, que no va a afectar al cálculo de la probabilidad a posteriori de B, pues el valor concreto que tome este vector "se pierde" al aplicar la normalización.

Siguiendo con la propagación de mensajes tenemos que $\lambda_C(a) = \sum_c [\lambda(c) \cdot \sum_b P(c|a,b) \cdot \pi_C(b)] = \sum_b [\sum_c P(c|a,b)] \cdot \pi_C(b) = (1\ 1)$, lo cual demuestra que la evidencia $\neg d$ no se propaga hasta A.

El último mensaje que se propaga entre nodos es $\pi_D(b) = \pi(b) \cdot \lambda_C(b) = \pi(b) = P(b) = (0'1\ 0'9)$. Nótese que el orden en que hemos calculado los mensajes es el siguiente: $\pi_C(a)$, $\lambda_D(b)$, $\pi_C(b)$, $\lambda_C(b)$, $\lambda_C(a)$ y $\pi_D(b)$.

Por cierto, observe que $\lambda_C(b) = (1\ 1)$ ha conducido a $\pi_D(b) = \pi(b) = P(b)$, que a su vez implica que $\pi(d) = \sum_b P(d|b) \cdot \pi_D(b) = \sum_b P(d|b) \cdot P(b) = P(d)$; es decir, $\pi(d)$ coincide con la probabilidad a priori de D, lo cual demuestra que la evidencia $+a$ no se ha propagado hasta D. Visto de forma más general, $\lambda(c) = (1\ 1)$ implica que $\lambda_C(b)$ y $\lambda_C(a)$ son vectores constantes que no propagan evidencia, y esto significa que cuando no hay información sobre C ni por debajo de C el camino A–C–B está desactivado, de modo que ni la evidencia $+a$ se propaga hasta B y D ni la evidencia $\neg d$ se propaga hasta A.

Finalmente, vamos a calcular los vectores $\pi()$ y $\lambda()$ que nos faltan, con el fin de poder aplicar la ecuación (3.72) a cada nodo y calcular así su probabilidad a posteriori. Para el nodo A, $\lambda(a) = \lambda_C(a) = (1\ 1)$, porque sólo tiene un hijo, C, que no aporta ninguna evidencia; por tanto, $P^*(a) = \alpha\, \pi(a)\, \lambda(a) = \alpha\, \pi(a) = (1\ 0)$; es decir, $P^*(+a) = 1$, como debe ser, pues $+a$ forma parte de la evidencia. Para B, $\lambda(b) = \lambda_C(\neg b)\, \lambda_D(\neg b) = (1\ 1) \cdot (0'2\ 1) = (0'2\ 1)$ y $P^*(b) = \alpha\, \pi(b)\, \lambda(b) = \alpha \cdot (0'1\ 0'9) \cdot (0'2\ 1) = (0'022\ 0'978)$. Para C, $\pi(c) = \sum_a \sum_b P(c|a,b)\, \pi_C(a)\, \pi_C(b) = (0'198\ 0'722)$ y $P^*(c) = \alpha \cdot (0'198\ 0'722) \cdot (1\ 1) = (0'215\ 0'785)$. Por último, para D, $\pi(d) = \sum_b P(d|b) \cdot \pi_D(b) = (0'08\ 0'92)$ y $P^*(d) = \alpha\, \pi(d)\, \lambda(d) = \alpha \cdot (0'08\ 0'92) \cdot (0\ 1) = (0\ 1)$, lo cual también era de esperar, porque $\neg d$ forma parte de la evidencia. □

Ejemplo 3.13 Dada la misma red del ejemplo anterior, calcular $P(+a|+c,\neg d)$ y $P(+b|+c,\neg d)$.

Solución. En este caso la evidencia es $\mathbf{e} = \{+c, \neg d\}$. Por tanto, los mensajes de inicialización son: $\pi(a) = P(a) = (0'3\ 0'7)$, $\pi(b) = P(b) = (0'1\ 0'9)$, $\lambda(c) = (1\ 0)$ y $\lambda(d) = (0\ 1)$. Ahora hay que propagar la evidencia, y vamos a ver cómo se haría en cada una de las dos implementaciones que conocemos: la recursiva y la distribuida.

En la **implementación recursiva**, podemos tomar A como nodo pivote, ya que nos interesa calcular $P^*(a)$. Por eso A solicita el mensaje $\lambda_C(a)$ a su único vecino, C, que lo va a calcular así: $\lambda_C(a) = \sum_c [\lambda(c) \cdot \sum_b P(c|a,b) \cdot \pi_C(b)]$. Como C ya "conoce" $\lambda(c)$ y $P(c|a,b)$ sólo necesita conocer el mensaje $\pi_C(b)$, que lo solicita a su vecino B. Este mensaje se calcula así: $\pi_C(b) = \pi(b) \cdot \lambda_D(b)$; por tanto B necesita el mensaje $\lambda_D(b)$, que solicita a D. El nodo D realiza el cálculo $\lambda_D(b) = \sum_d \lambda(d) \cdot P(d|b) = (0'2\ 1)$ y responde a B con el mensaje solicitado. De este modo B ya puede calcular que $\pi_C(b) = (0'02\ 0'9)$ y pasar este mensaje a C, el cual calcula a su vez que $\lambda_C(a) = (0'198\ 0'096)$ y se lo pasa a A. Éste era el único mensaje que A estaba esperando, y por tanto ya puede calcular que $\lambda(a) = \lambda_C(a) = (0'198\ 0'096)$ y $P^*(a) = \alpha\, \pi(a)\, \lambda(a) = (0'4692\ 0'5308)$. Por tanto, la respuesta a la primera pregunta del ejercicio es $P(+a|+c,\neg d) = 0'4692$.

Vamos a responder ahora a la segunda pregunta del enunciado. Para ello tomamos B como nodo pivote. Este nodo ya "conoce" su $\pi(b)$ y debe calcular su $\lambda(b)$, que es $\lambda(b) = \lambda_C(b)\lambda_D(b)$. Por eso B debe solicitar a C el mensaje $\lambda_C(b)$, que es $\lambda_C(b) = \sum_c [\lambda(c) \cdot \sum_a P(c|a,b) \cdot \pi_C(a)]$.

Como C ya "conoce" $\lambda(c)$ y $P(c|a,b)$ sólo necesita el mensaje $\pi_C(a)$, que solicita a A. Dado que A no tiene más hijos que C, este mensaje es simplemente $\pi_C(a) = \pi(a) = (0'3\ 0'7)$, con lo cual C puede calcular que $\lambda_C(b) = (0'48\ 0'13)$. Como B "recuerda" que $\lambda_D(b) = (0'2\ 1)$ —lo solicitó a D cuando A era el nodo pivote— ya puede calcular que $\lambda(b) = (0'096\ 0'13)$ y $P^*(b) = \alpha\,\pi(b)\,\lambda(b) = (0'0758\ 0'9242)$. Por tanto, la respuesta a la segunda pregunta del enunciado es $P(+b|+c,\neg d) = 0'0758$.

La otra forma de resolver el problema es mediante una **implementación distribuida**, en la que, como hemos visto ya, no hay un nodo pivote que solicite los mensajes, sino que cada nodo "sabe" cuántos mensajes ha recibido y, en función de ello, "decide" qué mensajes puede enviar. El valor numérico de cada mensaje depende de la evidencia, pero el orden en que se calculan los mensajes es siempre el mismo, y por tanto es el mismo que en el ejemplo anterior: primero A y D envían los mensajes $\pi_C(a) = (0'3\ 0'7)$ y $\lambda_D(b) = (0'2\ 1)$, respectivamente; luego C y B intercambian los mensajes $\lambda_C(b) = (0'48\ 0'13)$ y $\pi_C(b) = (0'02\ 0'9)$, y finalmente estos dos nodos envían los mensajes $\lambda_C(a) = (0'198\ 0'096)$ y $\pi_D(b) = (0'048\ 0'117)$. Una vez concluida la propagación de mensajes entre nodos, cada nodo puede calcular su π, su λ y su probabilidad a posteriori: $\lambda(a) = (0'198\ 0'096)$, $P^*(a) = (0'4692\ 0'5308)$, $\lambda(b) = (0'096\ 0'13)$, $P^*(b) = (0'0758\ 0'9242)$, $\pi(c) = (0'1266\ 0'7934)$, $P^*(c) = (1\ 0)$, $\pi(d) = (0'0384\ 0'0096)$ y $P^*(d) = (0\ 1)$.

Observe que en ambas implementaciones la inicialización y el modo de computar cada mensaje es el mismo; lo único que varía es el mecanismo de control que decide en qué momento se calcula cada mensaje. Podríamos decir que en la implementación distribuida hay un nodo pivote que toma la iniciativa y solicita a sus vecinos los mensajes que necesita, y así sucesivamente, mientras que en la implementación distribuida cada nodo "decide" de forma autónoma qué mensajes debe enviar.

Ejercicio 3.14 Dada la evidencia $\neg c$, calcular las probabilidades a posteriori de las otras tres variables.

Solución. $P(+a|\neg c) = 0'2623$, $P(+b|\neg c) = 0'0623$, $P(+d|\neg c) = 0'0498$.

3.4 La puerta OR/MAX

3.4.1 La puerta OR binaria

Hemos visto que, en el caso general, la probabilidad condicional viene dada por una tabla, tal como la que aparece en la página 3.1. Por tanto, el número de parámetros requerido para una familia crece exponencialmente con el número de padres. Esto conlleva varios inconvenientes. El más grave es la obtención de dichos parámetros: si obtenemos los resultados a partir de una base de datos, necesitamos gran cantidad de casos para que los parámetros obtenidos sean fiables; si la ausencia de una base de datos nos obliga a recurrir a la estimación subjetiva de un experto humano, resultará muy complicado para él responder a tantísimas preguntas correspondientes a una casuística compleja: "¿Cuál es la probabilidad de que el paciente presente fiebre dado que tenga paludismo, neumonía y apendicitis y no tenga amigdalitis, ni meningitis, ni etc., etc.?"

El segundo problema que plantea el modelo general, una vez obtenidos los parámetros, es la cantidad de espacio de almacenamiento que requiere cuando el número de padres es grande (por ejemplo, para un nodo binario con 10 padres binarios, la tabla de probabilidad condicional tiene $2^{1+10} = 2.048$ parámetros). Y por último, otro grave inconveniente es que

el tiempo de computación para la propagación de evidencia crece también exponencialmente con el número de padres de la familia considerada.

Por estas razones, es conveniente buscar modelos simplificados de interacción causal que simplifiquen la construcción de RR.BB. y la computación de la probabilidad. Pearl [45] los llama modelos *canónicos* porque son aplicables a numerosos campos, no son soluciones *ad hoc* para resolver un problema concreto de un dominio particular. Los más famosos entre ellos son las puertas OR y MAX probabilistas, que suponen una generalización de los correspondientes modelos deterministas.

En la puerta OR probabilista (*noisy OR-gate* [45, sec. 4.3.2]) se supone que cada causa U_i actúa para producir el efecto X, pero existe un *inhibidor* I_i que bloquea la influencia; es como si U_i estuviera inactiva. Por tanto, el parámetro fundamental es la probabilidad de que actúe el inhibidor (q_i) o bien su parámetro complementario, $c_i = 1 - q_i$, la probabilidad de que la causa U_i actuando en ausencia de otras causas llegue a producir X:

$$P(+x|+u_i, \neg u_{j\,[j\neq i]}) = c_i = 1 - q_i$$

Tenemos así la probabilidad de X en el caso de que haya una única causa presente y las demás están ausentes. Para hallar la probabilidad de X en el caso de que haya más de una causa presente, se introduce la hipótesis de que X sólo está ausente cuando todas las causas están ausentes o cuando para cada causa U_i que está presente ha actuado el correspondiente inhibidor I_i. Se supone que no sólo las causas sino también los inhibidores actúan independientemente, lo cual implica la independencia en sentido probabilista. En consecuencia,

$$P(\neg x|\bar{u}) = \prod_{i\in T_U} q_i$$

donde T_U indica el subconjunto de las causas de X que están presentes ($T_U \subset \bar{U}$).

A partir de aquí podemos construir la tabla $P(x|\bar{u})$ y aplicar el algoritmo de propagación general desarrollado en la sección anterior. Pero así habríamos resuelto uno sólo de los inconvenientes anteriores (el de la obtención de los parámetros), pues ya vimos que la complejidad de este algoritmo crecía exponencialmente con el número de padres. Por fortuna, existen expresiones para la puerta OR que llevan a un tiempo de propagación proporcional al tamaño de la familia. Dichas expresiones se encuentran en [45]. Nosotros, en vez de deducirlas aquí, las presentaremos como un caso particular de las correspondientes a la puerta MAX que vamos a estudiar a continuación.

3.4.2 Definición de la puerta MAX

Existe una generalización de la puerta OR binaria, que fue propuesta por Max Henrion [30] como modelo para la obtención del conocimiento; el nombre de "puerta MAX", su formulación matemática y los algoritmos de propagación que discutimos a continuación fueron publicados por primera vez en [13].

Para llegar a una formulación matemática del modelo es necesario introducir previamente el siguiente concepto:

Definición 3.15 (Variable graduada) Es la variable X que puede estar ausente o presente con g_X grados de intensidad. Tiene por tanto g_X+1 valores posibles, a los que asignaremos enteros tales que $X=0$ significa "ausencia de X" y los números sucesivos indican grados de mayor intensidad.

Ejemplo 3.16 Supongamos que la variable X =Neumonía puede estar ausente o presente con tres grados de intensidad (g_X =3): leve, moderada o severa. Entonces $X = 0$ significa "el paciente no tiene neumonía", $X = 1$ significa "el paciente tiene neumonía leve", etc. □

Observe que el concepto de *graduada* no es sinónimo de *multivaluada*. De hecho, son dos conceptos independientes: por un lado, no todas las variables multivaluadas representan distintos grados de intensidad, y por otro lado, hay variables binarias graduadas, como son las que hemos visto hasta ahora de tipo presente/ausente o positivo/negativo, cuyos valores representábamos por $+x$ y $\neg x$. (La definición de variable graduada nos dice que a $\neg x$ le corresponde el valor 0 y a $+x$ el valor 1.) Más aún, *las variables que intervienen en la puerta OR binaria son siempre variables graduadas*, pues no tiene sentido plantear dicho modelo para variables no graduadas, tales como el sexo.

El modelo de interacción de las puertas OR/MAX es bastante general; aquí lo vamos a definir en el contexto de las RR.BB., aunque sería aplicable a otros métodos de tratamiento de la incertidumbre. Por simplificar la escritura, llamaremos $\tilde{U}_i$ al conjunto de todas las causas de X excluida U_i:

$$\tilde{U}_i \equiv \bar{U} \setminus U_i$$

Definición 3.17 (Puerta OR/MAX) En una red bayesiana, dada una variable graduada X con n padres $U_1, \ldots, U_n$ (también variables graduadas), decimos que interaccionan mediante una *puerta MAX* cuando se cumplen las dos condiciones siguientes:

$$1. \quad P(X = 0|\bar{U} = 0) = 1 \tag{3.81}$$

$$2. \quad P(X \leq x|\bar{u}) = \prod_i P(X \leq x|U_i = u_i, \tilde{U}_i = 0) \tag{3.82}$$

Si X y las U_i son todas binarias, se dice que interactúan mediante una *puerta OR*.

Podemos utilizar la notación $x^0 \equiv$"$X = 0$" para expresar ambas condiciones en forma abreviada como

$$1. \quad P(x^0|\bar{u}^0) = 1 \tag{3.83}$$

$$2. \quad P(X \leq x|\bar{u}) = \prod_i P(X \leq x|u_i, \tilde{u}_i^0) \tag{3.84}$$

Intentaremos ahora explicar el significado de esta definición. La primera condición es fácil de interpretar: significa que, si todas las causas que pueden producir X están ausentes, entonces tenemos la seguridad de que también X estará ausente. Más adelante relajaremos esta restricción.

La segunda condición (ec. (3.82)) nos dice que $X \leq x$ sólo cuando ninguna de las causas U_i (actuando como si las demás causas estuvieran ausentes) ha elevado X a un grado superior a x. Dicho con otras palabras, el grado que alcanza X es el *máximo* de los grados producidos por las causas actuando independientemente; ésta es la razón por la que se denomina "puerta MAX". Al igual que en la puerta OR binaria, el resultado es el máximo de los valores de las entradas; esta coincidencia era de esperar, pues el modelo graduado es tan sólo una generalización del caso binario.

La importancia de esta definición es que permite calcular todos los valores de $P(x|\bar{u})$ a partir de un reducido número de parámetros. Para la familia X, serán las probabilidades X

condicionadas a que una sola de las causas esté presente:

$$c_{X=x}^{U_i=u_i} \equiv P(X=x|\, U_i=u_i, \tilde{U}_i = 0) \tag{3.85}$$

que podemos escribir en forma abreviada como

$$c_x^{u_i} \equiv P(x|u_i, \tilde{u}_i^0) \tag{3.86}$$

En principio, el número de parámetros para el enlace U_iX es $(g_{U_i}+1)\cdot(g_X+1)$. Sin embargo, la suma de las probabilidades debe ser la unidad y, en consecuencia,

$$c_{x^0}^{u_i} = 1 - \sum_{x=1}^{g_X} c_x^{u_i}. \tag{3.87}$$

Por otra parte, la primera condición de la definición de la puerta OR (ec. (3.81)) es equivalente a decir que

$$c_x^{u_i^0} = \begin{cases} 1 & \text{para } x = 0 \\ 0 & \text{para } x \neq 0\,. \end{cases} \tag{3.88}$$

Por tanto, sólo se necesitan $g_{u_i}\cdot g_X$ parámetros para este enlace.

Ejemplo 3.18 Supongamos que tenemos una porción de red representada por la figura 3.11, y que cada una de las tres variables puede tomar los siguientes valores:

$$\begin{aligned} U_1 &= \text{Neumonía} \rightarrow \{\text{ausente, leve, moderada, severa}\} \\ U_2 &= \text{Paludismo} \rightarrow \{\text{ausente, presente}\} \\ X &= \text{Fiebre} \rightarrow \{\text{ausente, leve, elevada}\} \end{aligned}$$

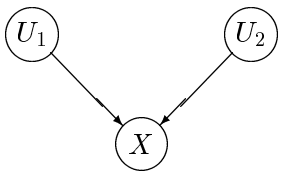

Figura 3.11: Ejemplo de puerta MAX.

Vemos que $g_{U_1}=3$, $g_{U_2}=1$ y $g_X=2$. En el modelo general, para esta familia necesitaríamos una tabla con 16 parámetros (hay 4×2=8 combianaciones posibles de u_1 y u_2; para cada una de ellas deberíamos dar tres valores $P(x|u_1,u_2)$, pero como la suma de los tres debe ser la unidad, sólo hace falta dar dos, de modo que necesitamos en total 8×2=16 parámetros). Sin embargo, para la puerta MAX, indicaremos los valores de $c_x^{u_1}$ y $c_x^{u_2}$, tal como muestran las tablas 3.2 y 3.3. En ellas se observa que el número de parámetros se ha reducido a la mitad. Si hubiéramos tenido más causas en vez de sólo dos, el ahorro habría sido mayor. A partir de estas dos pequeñas tablas, aplicando los axiomas anteriores podemos construir la tabla $P(x|u_1,u_2)$ completa necesaria para aplicar el algoritmo general. Sin embargo, existe una solución mucho más eficiente, que evita tener que calcular dicha tabla, como mostraremos en la próxima sección. □

$X \setminus U_1$	leve	moderada	severa
leve	0'50	0'40	0'20
elevada	0'20	0'50	0'80

Tabla 3.2: Parámetros $c_x^{u_1}$.

$X \setminus U_2$	presente
leve	0'20
elevada	0'75

Tabla 3.3: Parámetros $c_x^{u_2}$.

Causas no explícitas

Si en el modelo general tomamos $P(+x|\neg u_1, \neg u_2) > 0$, esto significa que X puede estar presente incluso cuando U_1 y U_2 están ausentes. Sin embargo, en la puerta OR/MAX, la propiedad (3.81) nos dice que cuando todas las causas de X están ausentes, sabemos con certeza que X estará ausente.

La propiedad en sí es razonable, pero existe el problema de que en la práctica es imposible considerar explícitamente todas las causas, pues éstas pueden ser muy numerosas e incluso muchas de ellas serán desconocidas; esto se ve especialmente claro en el caso de la medicina. La cuestión es importante aunque, afortunadamente, tiene una solución muy sencilla: para cada nodo X incluiremos un nodo X^* que agrupe todas las causas que no aparezcan explícitamente en el modelo.

Podemos suponer que el valor de este nodo siempre es presente ($\pi(+x^*) = 1$) y que su eficacia para producir X viene dada por los parámetros $c_x^{x^*}$ (de forma abreviada, c_x^*); en caso de que X sea una variable binaria, basta conocer un solo número, c_{+x}^*, pues $c_{\neg x}^*$ será siempre 0.

Al desarrollar el algoritmo de propagación para la puerta OR/MAX, veremos que el impacto de cada causa U_i se traduce en una $Q_{U_i}(x)$, y todas éstas se combinan de acuerdo con la ecuación (3.93). Por tanto, podemos aplicar, la propiedad asociativa del producto y agrupar varias causas en una sola sin violar los principios axiomáticos de las redes bayesianas. Lo que queremos decir es que está matemáticamente justificado incluir las causas no explícitas en un solo nodo y asignar al enlace correspondiente unos valores c_x^* que combinan los parámetros de todas ellas como si se tratara de una sola causa.

3.4.3 Algoritmo de propagación

Hemos resuelto ya los dos primeros problemas que presentaba el modelo general, pues ya no necesitamos obtener ni almacenar un número exponencial de parámetros por familia. Veamos a continuación cómo podemos resolver el tercero, es decir, cómo podemos realizar eficientemente la propagación de evidencia. Empezamos introduciendo la siguiente definición:

$$Q(x) \equiv P(X \leq x, \mathbf{e}_X^+) \tag{3.89}$$

Es fácil obtener $Q(x)$ a partir de $\pi(x)$

$$Q(x) = \sum_{x'=0}^{x} \pi(x') \tag{3.90}$$

y viceversa

$$\pi(x) = \begin{cases} Q(x) - Q(x-1) & \text{para } x \neq 0 \\ Q(0) & \text{para } x = 0 \end{cases} \tag{3.91}$$

Queremos encontrar ahora un algoritmo eficiente para calcular $Q(x)$. Aplicando las ecuaciones (3.83) y (3.73), podemos escribir

$$\begin{aligned} Q(x) &= \sum_{\bar{u}} P(X \leq x|\bar{u}) \quad P(\bar{u}, \mathbf{e}^+_X) \\ &= \sum_{\bar{u}} \left[\prod_i P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i|\mathbf{e}^+_{U_iX}) \right] \end{aligned}$$

En esta expresión podemos invertir el orden del productorio y de los sumatorios. En efecto,

$$\begin{aligned} &\sum_{u_1} \left[\prod_{i=1}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i, \mathbf{e}^+_{U_iX}) \right] \\ &= \left[\sum_{u_1} P(X \leq x|u_1, \tilde{u}_1^0) \quad P(u_1, \mathbf{e}^+_{U_1X}) \right] \cdot \prod_{i=2}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i, \mathbf{e}^+_{U_iX}) \\ &= P(X \leq x,\, \mathbf{e}^+_{U_1X},\, \tilde{u}_1^0) \prod_{i=2}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i|\mathbf{e}^+_{U_iX}) \\ &= Q_{U_1} \cdot \prod_{i=2}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i|\mathbf{e}^+_{U_iX}) \end{aligned}$$

donde hemos introducido la definición

$$Q_{U_i}(x) \equiv P(X \leq x, \mathbf{e}^+_{U_iX}, \tilde{u}_i^0) \tag{3.92}$$

que es la probabilidad de $X = x$ considerando toda la evidencia por encima del enlace U_iX, en caso que todas las demás causas de X estuvieran ausentes.

Sustituyendo este resultado en la expresión de $Q(x)$ tenemos

$$Q(x) = Q_{U_1}(x) \cdot \sum_{u_2,\ldots,u_n} \left[\prod_{i=2}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i|\mathbf{e}^+_{U_iX}) \right]$$

y repitiendo la misma operación n veces llegamos a

$$Q(x) = \prod_i Q_{U_i}(x) \tag{3.93}$$

Lo que necesitamos ahora es una fórmula sencilla para calcular $Q_{U_i}(x)$. Para ello, definimos un nuevo conjunto de parámetros $C_x^{u_i}$:

$$C_x^{u_i} \equiv P(X \leq x|u_i, \tilde{u}_i^0) \tag{3.94}$$

que podemos calcular a partir de las $c_x^{u_i}$, según las ecuaciones (3.86) y (3.87):

$$C_x^{u_i} = \sum_{x'=0}^{x} c_x^{u_i} = c_{x^0}^{u_i} + \sum_{x'=1}^{x} c_x^{u_i} = 1 - \sum_{x'=1}^{g_X} c_x^{u_i} + \sum_{x'=1}^{x} c_x^{u_i}$$
$$= 1 - \sum_{x'=x+1}^{g_x} c_x^{u_i} \tag{3.95}$$

Estos nuevos parámetros pueden ser almacenados junto con la descripción de la red (para ahorrar tiempo de computación) o calculados cuando se los necesita, aunque también es posible definir la red a partir de las $C_x^{u_i}$ en lugar de las $c_x^{u_i}$.

Desde aquí, el cálculo de $Q_{U_i}(x)$ es inmediato:

$$Q_{U_i}(x) = \sum_{u_i} P(X \leq x | u_i, \tilde{u}_1^0) \quad P(u_i, \mathbf{e}^+_{U_i X})$$
$$= \sum_{u_i} C_x^{u_i} \quad \pi_X(u_i) \tag{3.96}$$
$$= \sum_{u_i} \pi_X(u_i) \left[1 - \sum_{x'=x+1}^{g_X} c_x^{u_i} \right] \tag{3.97}$$

En el caso de que tengamos además unas c_x^* correspondientes a las causas no explícitas en el modelo, podemos manejarlas como si se tratara de una causa similar a las demás y calcular la respectiva $Q^*(X)$, que deberá incluirse en el productorio de la ecuación (3.93). El tratamiento de las causas no explícitas es, por tanto, muy sencillo.

Hemos resuelto ya la primera parte del problema: cómo calcular $\pi(x)$ para la familia X en tiempo proporcional al número de padres. También el cálculo de $\lambda(x)$ y el de $\pi_X(u_i)$ o $\lambda_{Y_j}(x)$ están resueltos, pues podemos aplicar las ecuaciones (3.77) y (3.75), ya que en ellas no aparece $P(x|\bar{u})$ y por tanto no varían al pasar del caso general a la puerta OR/MAX. Lo que nos falta por resolver es cómo calcular $\lambda_Y(u_i)$, es decir, el mensaje que X envía a cada uno de sus padres.

La ecuación (3.76) puede escribirse para la familia X como

$$\lambda_X(u_i) = \sum_x \left[\lambda(x) \sum_{\tilde{u}_i} P(x|\bar{u}) \prod_{j \neq i} \pi_X(u_j) \right] \tag{3.98}$$

Observe que, dentro de esta expresión, el valor de u_i en $\bar{u} = (u_1, \ldots, u_n)$ está fijo (depende de qué $\lambda_X(u_i)$ estamos calculando), mientras que el valor de las demás variables u_j va cambiando según indica el sumatorio.

Ahora bien, una forma de fijar el valor u_i' para la variable U_i es asignarle un vector $\pi(u_i)$ definido así:

$$[\pi(u_i')]_{U_i = u_i} = \begin{cases} 1 & \text{para } u_i = u_i' \\ 0 & \text{para } u_i \neq u_i' \end{cases} \tag{3.99}$$

puesto que entonces, según (3.72) y (3.75),

$$[P(u_i)]_{U_i = u_i'} = [\pi_X(u_i)]_{U_i = u_i'} = \begin{cases} 1 & \text{para } u_i = u_i' \\ 0 & \text{para } u_i \neq u_i' \end{cases}$$

y también

$$[P(x|\bar{u})]_{U_i=u'_i} = \sum_{u_i} P(x|\bar{u})\,[\pi(u_i)]_{U_i=u'_i}$$

Sustituyendo este resultado en la ecuación (3.98), tenemos

$$\lambda_X(u'_i) = \sum_x \left[\lambda(x) \sum_{\bar{u}} P(x|\bar{u}) \prod_j \pi_X(u_j)\right]_{U_i=u'_i}$$

o bien

$$\lambda_X(u_i) = \sum_x \lambda(x)\,[\pi(x)]_{U_i=u_i} \tag{3.100}$$

Aquí, $[\pi(x)]_{U_i=u_i}$ debe calcularse como hicimos anteriormente, es decir, con las ecuaciones (3.91) y (3.93), aunque ahora la ecuación (3.96) se simplifica para convertirse en

$$[Q_{U_i}(x)]_{U_i=u_i} = C_x^{u_i} \tag{3.101}$$

de acuerdo con el valor de $[\pi_X(u_i)]_{U_i=u'_i}$ indicado anteriormente.

Dicho con otras palabras, el algoritmo de la puerta OR/MAX puede expresarse así: para calcular $\pi(x)$, transformamos cada mensaje $\pi_X(u_i)$ en $Q_{U_i}(x)$, y los multiplicamos todos para obtener $Q(x)$, a partir del cual es muy sencillo obtener $\pi(x)$.

Cuando queremos calcular $\lambda_X(u_i)$ seguimos un procedimiento similar: para las causas U_j distintas de U_i tomamos las mismas $Q_{U_j}(x)$ que antes; para U_i, en cambio, tomamos la $Q_{U_i}(x)$ correspondiente al valor u_i según la ecuación (3.101), y repetimos —para cada valor de U_i— el mismo proceso que en el cálculo de $\pi(x)$.

3.4.4 Implementación distribuida

La implementación distribuida de la puerta OR/MAX es muy similar a la del caso general. Sin embargo, ahora los parámetros c_x^u (las tablas 3.2 y 3.3, por ejemplo) no son una característica del nodo X ni de esta familia en conjunto, sino que están asociados a cada enlace UX. Fíjese en la figura 3.12 y en la tabla 3.4 y observe dónde se almacenan las c_x^u. El nodo X debe saber solamente qué tipo de interacción debe aplicar: caso general o puerta OR/MAX; en el segundo caso, los parámetros se encontrarán almacenados en los enlaces correspondientes (salvo los parámetros c_x^*, correspondientes a las causas no explícitas, que se almacenarán en el propio nodo, para no tener que añadir un nodo que represente *OTRAS-CAUSAS-DE-X*). Comparando la figura 3.12 con la relativa al caso general (fig. 3.9, pág. 62), observamos que ahora los enlaces no son canales de información pasivos, sino procesadores activos que transforman cada mensaje $\pi_X(u)$ en $Q_U(x)$ y generan además $\lambda_X(u)$, liberando así al nodo X de algunas computaciones.

3.4.5 Semántica

Sería interesante desarrollar el ejemplo 3.18 para ver cómo funciona la propagación de evidencia en la puerta OR/MAX y comprobar que los resultados obtenidos matemáticamente coinciden con lo previsible mediante nuestro sentido común. Lamentablemente, desarrollar con detalle un solo ejemplo con los múltiples casos posibles nos ocuparía mucho más espacio de lo que podemos permitirnos. Por otra parte, puede encontrarse un ejemplo bien explicado

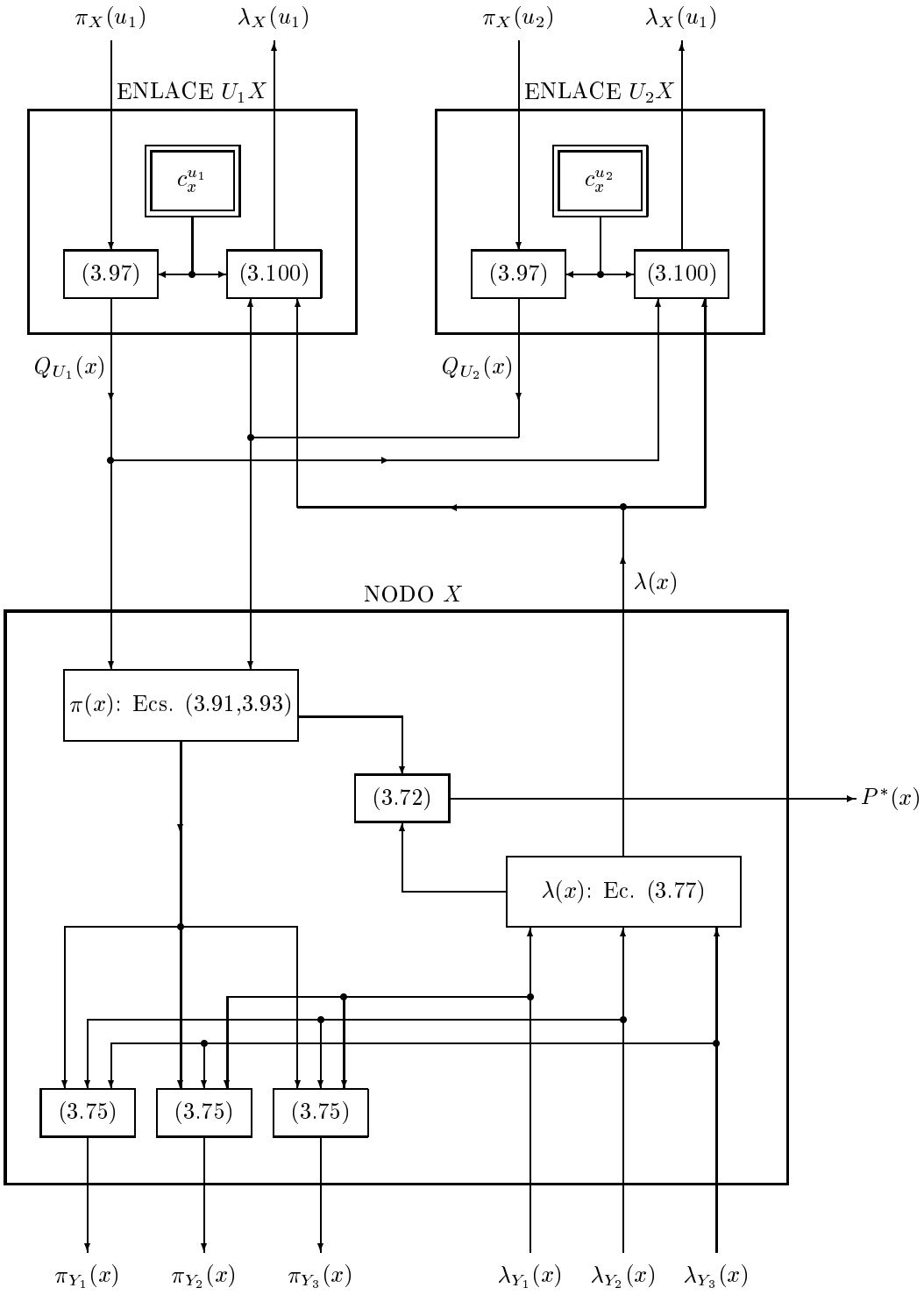

Figura 3.12: Computaciones realizadas en la puerta OR.

		Modelo general	Puerta OR
Nodo X	Almacena	$P(x\|u)$	c_x^*
	Recibe	$\pi_X(u_i)$, $\lambda_{Y_j}(x)$	$Q_{U_i}(X)$, $\lambda_{Y_j}(x)$
	Envía	$\lambda_X(u_i)$, $\pi_{Y_j}(x)$, $P^*(x)$	$\lambda(x)$, $\pi_{Y_j}(x)$, $P^*(x)$
Enlace U_iX	Almacena		$c_x^{u_i}$
	Recibe		$\pi_X(u_i)$, $\lambda(x)$, $Q_{U_{i'}}(x)$
	Envía		$Q_{U_i}(x)$, $\lambda_X(u_i)$

Tabla 3.4: Caso general y puerta OR.

en [45, sec. 4.3.2]; aunque allí se describe solamente la puerta OR, el tratamiento resulta muy similar al que pudiéramos realizar para la puerta MAX.

Lo que vamos a discutir en esta sección es la relación entre el modelo general y la puerta OR/MAX. En la sección 3.2.4 hablamos de la semántica de las redes bayesianas, refiriéndonos al modelo general. Allí vimos la relación entre los axiomas de independencia y los mecanismos causales percibidos intuitivamente. De igual modo, estudiar la semántica de la puerta OR/MAX consiste en establecer una relación entre los axiomas de la definición 3.17 y nuestros conceptos de causalidad. Esta cuestión fue abordada parcialmente al introducir dicha definición. En efecto, allí se mostró que la ecuación (3.81) significa que el efecto X está ausente cuando todas las causas que lo producen están ausentes, lo cual concuerda naturalmente con el sentido común, y la ecuación (3.82) significa que el grado que alcanza el efecto X es el máximo de los que producirían sus causas actuando independientemente.

Al igual que discutimos al hablar de la semántica de las redes bayesianas en general, podemos afirmar aquí que hay dos formas posibles de justificar la utilización de la puerta OR/MAX como modelo simplificado al construir nuestra red. La primera —la teórica— consiste en crear en nuestra mente un modelo de cómo actúan las causas a la hora de producir el efecto considerado. Por ejemplo, si suponemos que las diferentes causas de una enfermedad actúan independientemente, en el sentido de que, tal como dice la definición de puerta MAX, el grado más probable de la enfermedad es el máximo de los que producirían las causas, entonces estamos en condiciones de aplicar nuestro modelo simplificado; si no es así, debemos recurrir al modelo general.

La segunda forma de justificar nuestro modelo consiste en realizar estudios empíricos sobre un amplio número de casos y ver hasta qué punto la puerta OR/MAX puede considerarse como aproximación satisfactoria, y en esto pueden utilizarse dos criterios. Uno de ellos, el más estricto, consistiría en exigir que los resultados estadísticos para la familia X se ajustaran a los predichos por la expresión 3.82; el otro criterio, más flexible, se conformaría con que la red en su conjunto ofreciera diagnósticos acertados, dentro de ciertos límites, aunque las predicciones para la familia X no fueran completamente correctas.[10]

[10]Estos dos criterios pueden aplicarse también a sistemas expertos no bayesianos. Por ejemplo, en la eva-

Por último, al hablar de la semántica debemos insistir en la diferencia que existe entre el modelo general y la puerta OR/MAX. Si volvemos al ejemplo 3.3, comprobamos que hay una interacción entre el país de origen y el tipo sanguíneo como *factores condicionantes* de la probabilidad de contraer paludismo. Sin embargo, en el ejemplo 3.18, tenemos dos *causas*, neumonía y paludismo, cada una de las cuales por sí misma es capaz de producir fiebre, interactuando mediante una puerta MAX. Por tanto, podemos afirmar que la puerta OR/MAX refleja mejor el concepto intuitivo de causalidad que utilizamos en nuestra vida cotidiana. Así, cuando decimos que "A es una **causa** de C" entendemos que "A produce o puede producir C". Nótese el contraste con el primer ejemplo, referido al caso general: los nodos País-de-origen y Tipo-sanguíneo son los padres del nodo Paludismo, pero nadie diría "el país de origen produce paludismo" y menos aún "el tipo de sangre produce paludismo", sino "el país de origen y el tipo de sangre son dos **factores condicionantes** que influyen en la probabilidad de contraer paludismo".

De esta diferencia se deduce una ventaja más de la puerta OR/MAX frente al modelo general, además de las que habíamos mencionado anteriormente: a la hora de generar *explicaciones lingüísticas*, si los nodos de la familia X interactúan mediante una puerta OR/MAX podemos decir "la causa que [con mayor probabilidad] ha producido X es U_i", "la presencia de U_i explica por qué se ha producido X, y por tanto ya no es necesario sospechar la presencia de U_j" o "al descartar U_i por dichas razones, aumenta nuestra sospecha de que la causa más probable de X es U_j". En el modelo general, no es posible —al menos no es fácil— generar este tipo de explicaciones a partir de una tabla de probabilidades.

3.5 Bibliografía recomendada

Dado que las redes bayesianas son un tema de gran actualidad, la bibliografía relevante es extensa y crece día a día. Entre los libros publicados destaca el de Judea Pearl [45], que es la obra de referencia principal. Otro libro que recomendamos encarecidamente es el editado por J. A. Gámez y J. M. Puerta, *Sistemas Expertos Probabilísticos* [22]; sus ventajas principales son: que cubre casi todos los aspectos de las redes bayesianas (algoritmos de propagación, aprendizaje automático, modelos temporales, aplicaciones médicas e industriales...), que está escrito con fines didácticos, que ha aparecido muy recientemente, por lo que contiene los resultados y las referencias bibliográficas más actuales, y, aunque ésta es una ventaja de orden secundario, que está escrito en castellano. Otros libros didácticos, aunque todos ellos con aportaciones originales, son el de Neapolitan [41], el de Castillo, Gutiérrez y Hadi [6] y el de Jensen [31]. Otro libro excelente, aunque algo más complejo, es el de Cowell, Dawid, Lauritzen y Spiegelhalter [10]. También la tesis doctoral de F. J. Díez Vegas [14], disponible en Internet, puede servir de introducción al tema. La aplicación de las redes bayesianas y los diagramas de influencia a la medicina está descrita en [33] y [16].

Marek Druzdzel y Javier Díez están escribiendo actualmente un artículo que explica detalladamente la puerta OR/MAX y otros modelos canónicos, como las puertas AND, MIN, XOR, etc. El lector interesado en la aplicación de redes bayesianas a problemas del mundo real podrá encontrar dicho artículo a partir de noviembre o diciembre de 2001 en la página `http://www.dia.uned.es/~fjdiez/public.html`.

luación de MYCIN [66] se tomó el segundo de ellos y se consideró que la realización del programa había sido un éxito. Sin embargo, con el criterio más estricto, se habría cuestionado la validez del programa, pues éste contiene numerosas inconsistencias, como explicamos en la sección 4.4.

Quien tenga acceso a la WWW puede encontrar numerosos enlaces de interés a partir de la página `http://www.dia.uned.es/~fjdiez/bayes/rbayes.html`. En particular, recomendamos al alumno que trabaje con alguno de los programas gratuitos para redes bayesianas que se indican en ella, entre los que se encuentra el programa Elvira, desarrollado conjuntamente por varias universidades españolas.