

Redes Bayesianas

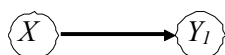
1 Presentación intuitiva

Antes de presentar formalmente la teoría matemática de las redes bayesianas, explicaremos mediante ejemplos sencillos¹ el significado intuitivo de los conceptos que después introduciremos

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

Ejemplo 1. La red bayesiana más simple

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 hasta la segunda.



Para concretar el ejemplo, supongamos que X representa paludismo e Y_1 representa gota gruesa, que es la prueba más habitual para detectar la presencia de dicha enfermedad.

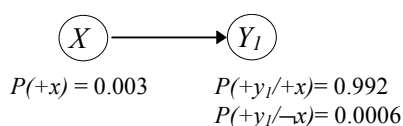
Cuando X sea una variable binaria, denotaremos por $+x$ la presencia de aquello a lo que representa y por $\neg x$ a su ausencia. Así, por ejemplo en este caso $+x$ significará “el paciente tiene paludismo” y $\neg x$ “el paciente no tiene paludismo”; $+y_1$ significará un resultado positivo del test de la gota gruesa y $\neg y_1$ un resultado negativo.

La información cuantitativa de una red bayesiana viene dada por:

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

Por tanto, en nuestro ejemplo, los datos que debemos conocer son $P(x)$ y $P(y_1/x)$.

Así, la red bayesiana completa sería:



Una red bayesiana consta de nodos, enlaces y parámetros

Los datos numéricos necesarios son:

- Probabilidades a priori de nodos sin padres
- Probabilidad condicionada de los demás

¹ Estos ejemplos están tomados de “Apuntes de razonamiento aproximado” de Francisco Javier Díez.

Veamos qué significado tienen en este caso estos valores:

- $P(+x) = 0.003$ indica que, a priori, un 0.3% de la población padece el paludismo. En medicina, esto se conoce como *prevalencia* de la enfermedad.
- $P(+y_1/+x) = 0.992$ indica que cuando hay paludismo, el test de la gota gruesa da positivo en el 99.2% de los casos. Esto se conoce como *sensibilidad* del test.
- $P(+y_1/-x) = 0.0006$ indica que, cuando no hay paludismo, el test de la gota gruesa da positivo en el 0.06% de los casos, y negativo en el 99.94%. A esta segunda probabilidad se la llama *especificidad* del test.

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

Alternativamente, se habla también de las *tasas de falsos positivos* (probabilidad de que el test de positivo si la persona no está enferma) y *tasas de falsos negativos* (probabilidad de test negativo cuando la persona está enferma).

Conociendo estos datos, podemos calcular:

- a) La probabilidad a priori de Y_1 ,

$$P(+y_1) = P(+y_1/+x) P(+x) + P(+y_1/-x) P(-x) = 0.00357.$$

$$P(-y_1) = P(-y_1/+x) P(+x) + P(-y_1/-x) P(-x) = 0.99643.$$

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

Supongamos que el test de la gota gruesa ha dado positivo. ¿Qué probabilidad hay ahora de que la persona padezca la enfermedad?. Si la prueba tuviese fiabilidad absoluta, esta probabilidad sería del 100%. Pero como existe la posibilidad de que haya habido un falso positivo, buscamos $P^*(+x) = P(+x/+y_1)$. Para calcularla, podemos aplicar el teorema de Bayes:

$$P^*(+x) = P(+x/+y_1) = \frac{P(+x) P(+y_1/+x)}{P(+y_1)} = \frac{0.003 \cdot 0.992}{0.00357} = 0.83263$$

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

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

$$P^*(-x) = P(-x/+y_1) = \frac{P(-x) P(+y_1/-x)}{P(+y_1)} = \frac{0.997 \cdot 0.0006}{0.00357} = 0.16737$$

Que, por supuesto, es la probabilidad complementaria.

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

Asociados a un test tenemos dos parámetros:

- Sensibilidad (probabilidad de resultado positivo si enfermo)
- Especificidad (probabilidad de resultado negativo si no enfermo)

Con la prevalencia, la sensibilidad y la especificidad, es posible calcular la probabilidad de que un paciente esté enfermo según el resultado de su test

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

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

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

Donde $\alpha = [P(y)]^{-1}$ y $\lambda_{Y_1}(x) = P(y/x)$.

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

Utilizando esta nueva expresión, podemos repetir los cálculos:

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

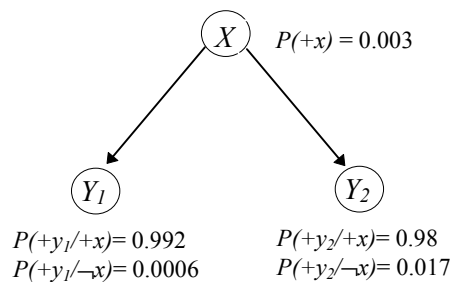
$$P^*(-x) = \alpha 0.997 0.0006 = 0.000598 \alpha.$$

Y normalizando obtenemos el mismo resultado que antes.

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

Ejemplo 2. Una red bayesiana con tres nodos

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



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

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

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

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

Cuando una variable se instancia o cambia su probabilidad, informa a su padre a través del paso de un λ -mensaje. En base a este mensaje, el padre actualiza su probabilidad

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

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

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

Para calcularla, usamos de nuevo el teorema de Bayes:

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

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

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

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

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

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

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

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

$$\lambda(+x) = \lambda_{y_1}(+x) \lambda_{y_2}(+x) = 0.97216$$

$$\lambda(-x) = \lambda_{y_1}(-x) \lambda_{y_2}(-x) = 0.0337.$$

Por tanto,

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

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

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

- e) En el caso en que tengamos un hallazgo a favor y otro en contra, podemos ponderar su influencia mediante estas mismas expresiones. Supongamos por ejemplo que el test de la gota gruesa ha dado

Las hipótesis de independencia condicional permiten en este ejemplo realizar el cálculo de las probabilidades a posteriori

negativo, pero que el paciente padece fiebre, esto es, $e = \{-y_1, +y_2\}$, y calculemos la probabilidad de que el paciente padezca fiebre.

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

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

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

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

$$P^*(\neg x) = 0.9986$$

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

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

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

Aplicando la hipótesis de independencia condicional y llamando

$$\pi_{y_1}(x) = P(x, y_2) = P(x) P(y_2/x)$$

$$\alpha = [P(y_2)]^{-1},$$

la expresión anterior nos queda:

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

Sustituyendo los valores numéricos de nuestro ejemplo, tenemos que,

$$\pi_{y_1}(+x) = P(+x)P(+y_2/+x) = 0.003 \cdot 0.98 = 0.00294$$

$$\pi_{y_1}(\neg x) = P(\neg x)P(+y_2/\neg x) = 0.997 \cdot 0.017 = 0.01695$$

Y, finalmente,

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

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

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

Cuando una variable se instancia o cambia su probabilidad, informa a su hijo a través del paso de un π -mensaje. En base a este mensaje, el hijo actualiza su probabilidad

probabilidad de x tras considerar la evidencia relativa a x *diferente* de Y_1 .

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

En las redes bayesianas, la información circula hacia arriba y hacia abajo, permitiendo hacer inferencias tanto abductivas como predictivas

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

2 Definición formal de red bayesiana

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

Definiciones previas

- **Arco.** Es un par ordenado (X, Y) . Esta definición de arco corresponde a lo que en otros lugares se denomina arco dirigido. En la representación gráfica, un arco (X, Y) viene dado por una flecha desde X hasta Y .
- **Grafo dirigido.** Es un par $G = (N, A)$ donde N es un conjunto de nodos y A un conjunto de arcos definidos sobre los nodos.
- **Grafo no dirigido.** Es un par $G = (N, A)$ donde N es un conjunto de nodos y A un conjunto de arcos no orientados (es decir, pares no ordenados (X, Y)) definidos sobre los nodos.
- **Camino.** Es una secuencia ordenada de nodos $(X_{i_1}, \dots, X_{i_r})$ tal que $\forall j = 1, \dots, r-1$, ó bien el arco $X_j \rightarrow X_{j+1} \in A$ o bien el arco $X_{j+1} \rightarrow X_j \in A$.
- **Camino dirigido.** Es una secuencia ordenada de nodos $(X_{i_1}, \dots, X_{i_r})$ tal que para todo $j = 1, \dots, r-1$ el arco $X_j \rightarrow X_{j+1} \in A$.
- **Ciclo:** es un camino no dirigido que empieza y termina en el mismo nodo X .
- **Grafo acíclico:** es un grafo que no contiene ciclos.
- **Padre.** X es un *padre* de Y si y sólo si existe un arco $X \rightarrow Y$. Se dice también que Y es **hijo** de X . Al conjunto de los padres de X se representa como $pa(X)$, y al de los hijos de X por $S(X)$.
- **Antepasado o ascendiente.** X es un *antepasado* o ascendiente de Z si y sólo si existe un camino dirigido de X a Z .
- **Conjunto ancestral** de un nodo X es un conjunto que contiene a X y a todos sus antepasados.

- **Descendiente.** Z es un *descendiente* de X si y sólo si X es un antepasado de Z . Al conjunto de los descendientes de X lo denotaremos por $de(X)$.
- **Variable proposicional** es una variable aleatoria que toma un conjunto exhaustivo y excluyente de valores. La denotaremos con letras mayúsculas, por ejemplo X , y a un valor cualquiera de la variable con la misma letra en minúscula, x .
- Dos variables X e Y son **independientes** si se tiene que $P(X/Y) = P(X)$. De esta definición se tiene una caracterización de la independencia que se puede utilizar como definición alternativa: X e Y son independientes si y sólo si $P(X,Y) = P(X) \cdot P(Y)$.
- Dos variables X e Y son **independientes** dado una tercera variable Z si se tiene que $P(X/Y,Z) = P(X/Y)$. De esta definición se tiene una caracterización de la independencia que se puede utilizar como definición alternativa: X e Y son independientes dado Z si y sólo si $P(X,Y/Z) = P(X/Z) \cdot P(Y/Z)$. También se dice que Z *separa condicionalmente* a X e Y .

Definición

Una red bayesiana es:

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

tales que:

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

Hipótesis de independencia condicional.

Un grafo acíclico conexo y dirigido $G = (V, E)$ y una distribución de probabilidad conjunta P definida sobre las variables del grafo se dice que cumplen las hipótesis de independencia condicional, si para toda variable X de V se tiene que el conjunto de los padres directos de X , que denotaremos por $pa(X)$ *separa condicionalmente* a X de todo otro nodo Y de la red que no sea X , ni sus descendientes ni sus padres.

$$\forall X \in V \text{ y } \forall Y \subset V - \{X \cup de(X) \cup pa(X)\} \text{ se tiene que } P(X/pa(X), Y) = P(X/pa(X)).$$

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

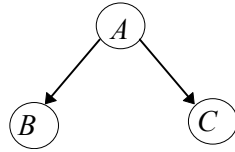
Vamos a hacer un ejemplo en el que demostraremos que una red es una red bayesiana.

Los nodos de una red bayesiana deben ser variables proposicionales (toman un conjunto exhaustivo y excluyente de valores)

Las hipótesis de independencia condicional establecen que cada nodo debe ser independiente de los otros nodos de la red (salvo sus descendientes) dados sus padres

Ejemplo 3. Comprobando si una red es bayesiana

Consideremos la red dada en la siguiente figura



$$P(a_1) = 0.3; \quad P(b_1/a_1) = 0.4 \quad P(b_1/a_2) = 0.2 \quad P(c_1/a_1) = 0.7 \quad P(c_1/a_2) = 0.6$$

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

$P(a_1, b_1, c_1) = 0.084$	$P(a_1, b_1, c_2) = 0.036$
$P(a_1, b_2, c_1) = 0.126$	$P(a_1, b_2, c_2) = 0.054$
$P(a_2, b_1, c_1) = 0.084$	$P(a_2, b_2, c_1) = 0.056$
$P(a_2, b_1, c_2) = 0.336$	$P(a_2, b_2, c_2) = 0.224$

¿Es esto una red bayesiana?

En este caso, la condición de independencia condicional se satisface si C y B son independientes dado su padre común A, es decir, queda reducida a probar que $P(B/A, C) = P(B/A)$. Nótese que esta red presenta la misma estructura que la del ejemplo 2, y la hipótesis que ahora tenemos que demostrar es la misma que tuvimos que hacer allí para poder seguir con los cálculos.

Para probar esto, tendríamos que ver que

$$\begin{aligned} P(b_1/c_1, a_1) &= P(b_1/c_2, a_1) = P(b_1/a_1) \\ P(b_1/c_1, a_2) &= P(b_1/c_2, a_2) = P(b_1/a_2) \end{aligned}$$

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

A modo de ejemplo vamos a comprobar una de ellas.

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

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

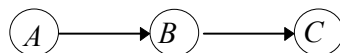
Para este tipo de estructura se dice que los enlaces convergen cola-con-cola en el nodo A. De este modo, conocer el valor del padre común (A) cierra la comunicación entre los hijos (B y C), es decir, una vez que se conoce el valor de A, conocer el valor que toma un hijo ya no aporta información sobre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

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

En la estructura cola-con-cola la comunicación entre los hijos está abierta, y se cierra al conocer el valor del padre común

Veamos otros ejemplos,

- Para la red ;



Habría que comprobar que B separa condicionalmente a A de C .

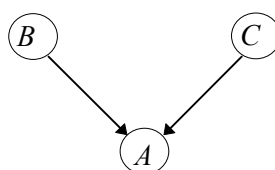
En este tipo de estructura se dice que los enlaces convergen *cola-con-cabeza* en B . Al conocer el valor de B , se *cierra la comunicación* entre el padre de B y el hijo de B .

En la estructura cola-con-cabeza, la comunicación entre los nodos raíz y hoja está abierta, y se cierra al conocer el valor nodo intermedio

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

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

- Para la red ;



Habría que comprobar que B es independiente de C .

En la estructura cabeza-con-cabeza la comunicación está cerrada, y se abre al conocer el valor del hijo común

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

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

El efecto explaining-away permite que conforme una de las posibles explicaciones cobra fuerza, las otras se vayan debilitando.

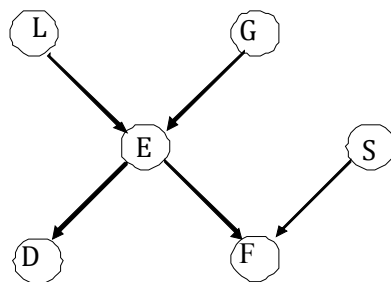
Como se puede ver, las comprobaciones de independencia necesarias dependen de la estructura de la red.

Ejemplo 4. La red Asia

Consideremos las siguientes variables binarias:

L = situación laboral
G = ganancias por inversiones
E = situación económica
S = salud
D = donaciones
F = felicidad.

Entre ellas existen las relaciones causales que se reflejan en esta red:



Especifica qué independencias condicionales deben cumplirse para que esta sea una red bayesiana.

Aplicando las hipótesis de independencia condicional, tendríamos que:

- D independiente de F, S, G, L dado E
- F independiente de D, L, G dado {E, S}

- E independiente de S dados {L, G}
- S independiente de L, G, E, D
- L independiente de G y S
- G independiente de L y S

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

Teorema (Factorización de la probabilidad)

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

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

Demostración:

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

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

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

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

Con lo que concluimos la demostración.

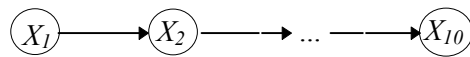
La importancia de este teorema es que nos permite describir una red bayesiana a partir de la probabilidad condicionada de cada nodo (o la probabilidad a priori en el caso de nodos sin padres) en lugar de dar la probabilidad conjunta, que,

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

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

Si se cumplen las condiciones de independencia condicional, a partir de las probabilidades condicionadas es posible calcular la distribución conjunta

Por ejemplo, para la red bayesiana dada por:



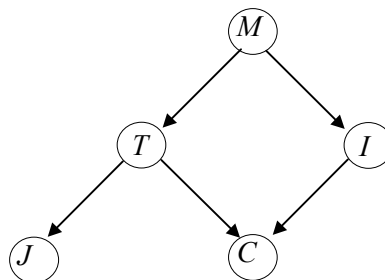
Suponiendo que todas las variables fuesen binarias, para dar la distribución conjunta habría que dar $2^{10}-1$ valores, sin embargo, si construimos la distribución conjunta a partir de los 19 valores necesarios para dar las condicionadas, tendremos además asegurado que se satisfacen las hipótesis de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

Ejemplo 5. *Independencias, distribuciones condicionadas y distribución conjunta*

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

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

Representamos dicha información en una siguiente red bayesiana:



¿Qué independencias implica la red?

- I es independiente de T, J dado M. Así por ejemplo dado metástasis, el incremento de calcio no depende de si hay o no tumor cerebral.
- T independiente de I dado M.
- C independiente de M, J dados {T, I}
- J independiente M, I, C dado T.

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

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

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

$$P(m_1, i_2, t_1, j_2, e_2) = P(m_1) \cdot P(i_2/m_1) \cdot P(t_1/m_1) \cdot P(c_2/i_2, t_1) \cdot P(j_2/t_1) = \\ = 0.2 \cdot 0.2 \cdot 0.2 \cdot 0.3 \cdot 0.2 = 0.00048.$$

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

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

Como vemos, una vez conocida la distribución conjunta es posible conocer la distribución de cualquier conjunto de variables, condicionadas o no condicionadas.

Sin embargo, este método de cálculo es computacionalmente muy costoso. Desde la aparición de las redes bayesianas, se han desarrollado varios algoritmos más eficientes de propagación de probabilidades. Aunque el procedimiento de cálculo exacto de las probabilidades es NP-duro (Cooper, 1987), hoy en día existen algoritmos exactos que se ejecutan en tiempo real y, para redes con estructuras muy complejas (elevado número de padres para cada nodo), algoritmos aproximados que proporcionan una buena estimación de las probabilidades en tiempos razonables. Además, se dispone de bibliotecas que permiten integrar estos algoritmos de forma sencilla en nuestras aplicaciones. En la sección 4 veremos en detalle el algoritmo exacto para redes con estructura de árbol, que es el caso más simple.

3 Modelado con redes bayesianas

Una vez hemos definido las redes bayesianas, vamos a aprender a modelar problemas de la vida real utilizando este enfoque. En esta sección utilizaremos el siguiente ejemplo.

Ejemplo 6. *El ejemplo del estornudo*

Una tarde, Juan va a visitar a sus amigos Pablo y Lara. De repente, comienza a estornudar. Juan piensa que se ha resfriado, hasta que observa que los muebles de la casa están arañados. Entonces, especula con la posibilidad de que sus amigos tengan un gato y sus estornudos se deban a una crisis de la alergia a los gatos que tiene diagnosticada.

Principalmente, los tipos de problemas que se suelen modelar con redes bayesianas son problemas de diagnóstico o problemas de predicción. El ejemplo de la alergia es un problema de diagnóstico, puesto que Juan intenta determinar la causa de sus estornudos.

3.1 Identificación de las variables

En primer lugar, es importante estudiar el dominio para tener el grado máximo de conocimiento y comprensión sobre el problema que vamos a modelar. En la mayoría de los casos reales, esto nos obligará a contar con expertos en el área, que deberán estar suficientemente interesados y motivados para que la colaboración tenga buenos frutos.

Una vez conocemos suficientemente el problema, el siguiente paso consiste en identificar las variables que son relevantes. Es importante centrarse sólo en aquellas variables que son de interés en el problema actual. Para ello, ayuda realizarse preguntas del tipo:

- ¿Cuál es la situación/problema que se plantea?
- ¿Qué posibles causas pueden explicar esta situación?
- ¿Qué otros factores pueden hacer que los problemas o causas ocurran, o impedir que ocurran?
- ¿De qué evidencia se dispone para soportar dichas causas, problemas o factores?

Es importante identificar las variables relevantes en el problema

Veamos cómo aplicar esto a nuestro ejemplo. En nuestro caso, el problema parece ser que Juan está estornudando. Las causas posibles son que se ha resfriado o que tiene rinitis. La rinitis puede estar causada porque sus amigos tienen un gato y Juan es alérgico a los gatos. La evidencia que sugiere que sus amigos tienen un gato es que algunos muebles tienen arañazos. La información relevante de la situación está contenida en las seis palabras subrayadas. Ejemplo de información irrelevante en este caso es que Juan y Pablo son amigos o que Juan está visitando a Pablo.

En el caso general del modelado de problemas de diagnóstico, hay ciertos tipos de variables susceptibles de ser agrupados en clases. Si se aborda el problema teniendo estas clases en mente, el proceso de modelado resulta más sencillo. Hablaremos por tanto de estas clases.

Variables objetivo

Estas variables se usan para modelar los objetos de interés, es decir, aquellos objetos sobre lo que nos gustaría razonar. Las variables objetivo suelen utilizarse para modelar fenómenos *latentes*, es decir, fenómenos que no son directamente observables. En el ejemplo del estornudo, Juan piensa en dos alternativas: o bien se ha Resfriado o bien tiene Alergia. Ambos son ejemplos de variables objetivo ya que Juan está interesado en saber más sobre ellas (el estado en el que están o los valores que tienen). En diagnóstico médico, las enfermedades serían modeladas como variables objetivo.

Variables de observación

Las variables de observación se usan para modelar las formas indirectas que tenemos de medir las variables objetivo. También se denominan *variables de evidencia*. En el ejemplo del estornudo, Juan piensa que está bien hasta que

empieza a estornudar. Sólo después de observarse a sí mismo estornudando se pregunta si está Resfriado. *Estornudar* sería una variable de observación. Otra podría ser *Arañazos*, porque Juan hace esa observación y la usa para razonar sobre la posibilidad de que exista un gato en casa (por el momento, no directamente observable). En el diagnóstico médico, los síntomas que muestra el paciente y los resultados de sus pruebas serían modeladas como observaciones. Algunas observaciones pueden ser obligatorias. Por ejemplo, en el diagnóstico médico un tipo específico de escáner puede ser requisito indispensable con el objetivo de detectar un posible cáncer.

Factores

Estas variables se usan para modelar los fenómenos que afectan a las variables objetivo. También se denominan *variables de contexto*. En el ejemplo del estornudo, la estación del año podría ser un factor que afecta al resfriado, pues es más probable que una persona se resfríe en invierno que en verano.

Los factores pueden dividirse en cuatro categorías, con respecto al tipo de influencia en las variables afectadas.

- *Promotores*. Si el factor promotor ocurre, la variable afectada será más probable (correlación positiva). Por ejemplo, fumar puede incrementar las probabilidades de tener un cáncer de pulmón.
- *Inhibidores*. Si el factor promotor ocurre, la variable afectada es menos probable (correlación negativa). Por ejemplo, practicar deporte puede disminuir las probabilidades de caer enfermo.
- *Requeridos*. Es indispensable que estos factores entren en acción para sea posible que ocurran las variables afectadas. Por ejemplo, para que una población específica de bacterias crezca se requiere que la temperatura esté por encima de un determinado nivel.
- *Preventivos*. Si el factor ocurre, la variable afectada no puede ocurrir. Por ejemplo, recibir la vacuna de la viruela a edades tempranas previene que se sufra esta enfermedad.
- *Auxiliares*. Son variables que se usan por conveniencia. Por ejemplo, para simplificar el proceso de modelado y especificación de parámetros.

Esta lista de tipos de variables no pretende ser exhaustiva, sino simplemente ilustrar diversos tipos existentes. La resumimos en la siguiente tabla.

Tipo de variable	Breve descripción
Objetivo	Modelan objetos de interés. No observables directamente.
Observación	Modelan la forma de medir variables objetivo. Pueden ser observadas directamente
Factor	Modelan fenómenos que afectan a otras variables del modelo.
Promotor	La variable afectada es más probable cuando están presentes.
Inhibidor	La variable afectada es menos probable cuando están presentes.
Requerido	Si no entra en acción, no ocurre la variable afectada.
Preventivo	Si entra en acción, no ocurre la variable afectada.
Auxiliares	Usadas por conveniencia (para simplificar el modelo)

Tabla A. Tipos de variables en modelado bayesiano

3.2 Estados y valores

Hasta ahora hemos definido varios tipos de variables con respecto al papel que juegan en el modelado de un problema. Pero también podemos dividir las según su escala de medición:

Variables cualitativas:

Son las variables que expresan distintas cualidades, características o modalidad. Cada modalidad que se presenta se denomina atributo o categoría y la medición consiste en una clasificación de dichos atributos. Las variables cualitativas pueden ser *dicotómicas* cuando sólo pueden tomar dos valores posibles como sí y no, hombre y mujer, o *politómicas*, cuando pueden adquirir tres o más valores.

Variables cuantitativas:

Son las variables que se expresan mediante cantidades numéricas. Las variables cuantitativas además pueden ser *discretas*, que presentan separaciones o interrupciones en la escala de valores que puede tomar (por ejemplo, el número de hijos), o *continuas*, que pueden adquirir cualquier valor dentro de un rango especificado (por ejemplo la edad).

A menudo conviene representar un fenómeno continuo en la naturaleza usando variables discretas. Para ello, las medidas continuas tienen que ser discretizadas. Esto puede hacerse proyectando la escala de valores continua en un conjunto finito de intervalos. Los valores que caigan en el mismo rango se considerarán como un mismo estado. Un ejemplo de discretización es modelar la variable temperatura con tres estados: bajo, medio, y alto.

La definición de variable proposicional tiene su importancia a la hora de modelar un problema con una red bayesiana, ya que deberemos tener en cuenta *que los nodos de la red son variables proposicionales y por tanto deben tomar un conjunto exhaustivo y excluyente de valores.*

De este modo, si por ejemplo estamos construyendo un sistema de diagnóstico médico en que las enfermedades posibles son *gripe*, *faringitis* y *alergia*, cada una de estas enfermedades será representada por una variable dicotómica diferente (que tomará valores *si/no*), ya que nada impide que un

paciente padezca dos o más enfermedades a la vez. Es decir, al no conformar las enfermedades un conjunto exhaustivo y excluyente de variables, cada una de ellas debe ser modelada como una variable dicotómica y no como valores de una única variable.

Sin embargo, si estamos construyendo un sistema de clasificación de animales en el que hemos representado todas las posibilidades (mamífero, ave, reptil, pez, etc), debemos introducir una única variable, cuyos estados serán las diferentes clases consideradas (ya que un animal no puede ser a la vez un mamífero y un reptil). A veces, si no se está completamente seguro de que el conjunto considerado es exhaustivo, se puede añadir un estado indeterminado “otro” de modo que se cumplan todas las condiciones para que cada nodo contenga una variable proposicional.

En el ejemplo del estornudo consideraremos una variable por cada dato identificado como relevante (dado que las variables no son incompatibles entre sí), y que todas las variables son dicotómicas, de forma que cada una de las variables tomará el valor “*presente*” o “*ausente*”. En problemas y aplicaciones más sofisticados se podría usar variables discretas (por ejemplo, la variable estornudo podría tomar los valores pocos, algunos, muchos) o variables continuas (por ejemplo, si tuviéramos una variable temperatura, podría tomar los valores que van desde 35,8 a 42,5). Pero evidentemente, cuanto más valores tomen las variables, más complicado será el modelo, así que a la hora de decidir los estados deberíamos tener presente qué grado de granularidad es realmente necesario en nuestra aplicación.

3.3 Estructura

Después de definir las variables, el siguiente paso en la construcción de un modelo es definir su estructura. Esto lo hacemos conectando variables con arcos (también llamados enlaces). Como hemos visto, en las redes bayesianas los arcos son dirigidos. Cambiar la dirección de un arco cambia su significado. La ausencia de un arco entre dos variables indica que no existen relaciones de dependencia directa entre ellas, sino a lo sumo a través de otras variables. En el ejemplo del estornudo las variables Estornudo y Gato no son directamente dependientes, por lo que no debería haber un arco que las una².

La presencia de un arco indica una *relación de influencia causal* entre dos variables. Por ejemplo, la *presencia* o *ausencia* de una enfermedad tiene influencia que el resultado de las pruebas sea *positivo* o *negativo* (Figura 2). En el ejemplo del estornudo, Gato tiene influencia causal en Arañazos, y por tanto debe existir un arco entre estas dos variables.

² A veces será difícil juzgar cual es la causa y cual el efecto. Por ejemplo, pensemos en el caso de una persona que no come y padece anorexia nerviosa. Podemos pensar que anorexia es la causa de que no coma, pero también que la anorexia ha venido provocada por no comer. En ese caso una opción es pensar lo que ocurrió antes en el tiempo, y modelarlo de forma que sea la causa mientras que la variable posterior será considerada un efecto.

A la hora de decidir si una entidad del mundo real debe ser una variable o un estado de una variable, debemos recordar que las variables de una red deben tomar un conjunto exhaustivo y excluyente de valores

No debemos introducir en el modelo mayor nivel de detalle del realmente necesario para nuestra aplicación

A la hora de definir los arcos, debemos reflejar en ellos las relaciones de *influencia causal* entre las variables.

Una alternativa a la dirección causal es la dirección de diagnóstico (Figura 2). La dirección de diagnóstico surge de aplicar reglas de diagnóstico del tipo “Si el paciente tiene tos, entonces el paciente tiene gripe”. En este caso, la variable de observación precede a la variable objetivo. Utilizar este tipo de relaciones para modelado en redes bayesianas no es en sí un error, pero la dificultad añadida que supone hace que frecuentemente se usen de un modo incorrecto.

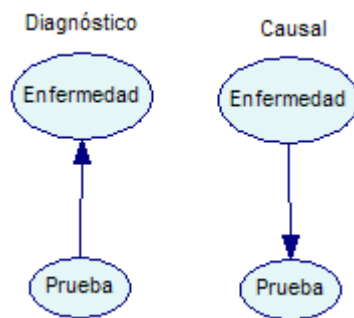


Figura 2. Posibles direcciones de los arcos.

El marco de trabajo de las RBs no impone que los arcos se construyan siempre en la dirección causal. Según Pearl, “los patrones de independencia plasmados en un grafo acíclico dirigido son típicos de organizaciones causales” (Pearl, 1998). También, Druzel y Simon explican que “Se pueden usar muchos modelos equivalentes para representar el mismo sistema ... pero se prefieren los modelos que utilicen las direcciones causales, ya que minimizan el número de arcos en el grafo, aumentando la claridad de los modelos y ofreciendo ventajas computacionales” (Druzel y Simon, 1993).

Utilizar relaciones causales conduce a modelos más sencillos de especificar y entender

Para ilustrar el motivo por el que se prefiere la dirección causal, vamos a construir una RB causal para el ejemplo de los estornudos. Una vez definida, calcularemos la red equivalente con los enlaces invertidos para mostrar que el modelo obtenido es bastante más difícil de interpretar.

Vamos ahora a desarrollar el modelo del ejemplo del estornudo. Para ello vamos a asignar a cada información que hemos considerado relevante en nuestro modelo una variable o nodo, y los vamos a ir incluyendo en una red que construiremos de modo incremental:

En este problema, la situación a la que queremos ofrecer una explicación es a los estornudos de Juan. Por ello, introducimos en la red un nodo que llamaremos Estornudo:

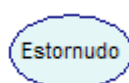


Figura 3. Primer nodo para la RB del estornudo.

Las causas posible del estornudo son Resfriado o Rinitis. La añadimos al modelo y ahora la red es:



Figura 4. Añadiendo causas a la RB del estornudo.

Entonces Juan observa Arañazos en los muebles:



Figura 5. Añadiendo fuentes de evidencia a la RB del estornudo.

Juan empieza a pensar que sus amigos pueden tener un Gato, lo que puede causar los Arañazos:



Figura 6. Añadiendo explicaciones a la RB del estornudo.

Lo que causa que Juan piense que él está estornudando podría ser debido a la Rinitis, causada por la Alergia y la presencia del Gato:

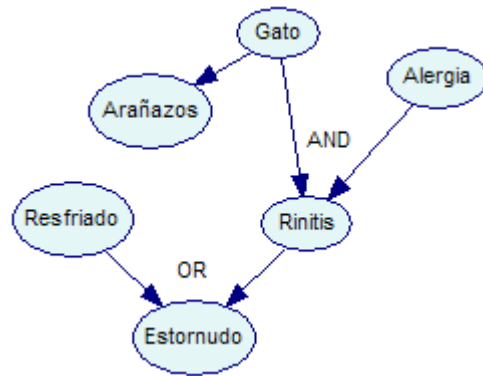


Figura 7. La RB final del problema del estornudo

De esta forma, hemos construido de modo incrementalmente el modelo de RB para el ejemplo del estornudo.

3.4 Parámetros

El último paso en el proceso de modelado es *especificar parámetros*. Como se explicó antes, basta con proporcionar las probabilidades a priori de los nodos raíz y las probabilidades condicionales del resto de los nodos

Existen varias alternativas para obtener los parámetros necesarios de una red:

- Especificación directa de los parámetros, normalmente contando con la ayuda de expertos. Este procedimiento es ciertamente costoso.
- Aprendizaje a partir de bases de datos, que obviamente depende de la existencia de dicha base de datos. De ser así, se tienen dos opciones; a) aprendizaje de los parámetros, si se dispone de la estructura; b) aprendizaje estructural, en el que es posible aprender tanto la estructura como los parámetros.; y
- Combinar especificación y aprendizaje. Por ejemplo, contar con expertos que nos ayuden a especificar la estructura, aprender los parámetros y disponer de expertos de nuevo para supervisar el modelo obtenido. Normalmente, esta última alternativa ofrece lo mejor de cada caso.

Usaremos el ejemplo del resfriado para ilustrar el proceso de especificación de parámetros. La notación que usaremos será la siguiente: el nombre de la variable indicará su presencia, y el nombre precedido de un símbolo ~ indicará su ausencia.

Para los nodos raíz, necesitamos proporcionar las probabilidades anteriores, que son las probabilidades a priori o en ausencia de información. Por ejemplo, necesitamos proporcionar la probabilidad del nodo Gato. Para ello, podemos pensar en la proporción de familias que tienen gato. Si disponemos información disponible sobre esto (por ejemplo, estadísticas) podemos usarlas, en otro caso, tendremos que hacer uso de una estimación

razonable. Vamos a considerar que un 20% de las familias tienen gato. Entonces la probabilidad anterior de que una familia tenga un gato es 0.2.

Para aquellos nodos que tienen padres, necesitamos proporcionar probabilidades condicionadas. Por ejemplo, necesitamos proporcionar la probabilidad de *estornudo* *dado* *resfriado* y *rinitis*. Para ello, necesitamos pensar en la clase de relación entre estas tres variables, que en este caso es una relación tipo OR, es decir, tanto el resfriado como la rinitis pueden causar el estornudo. Si tenemos ambas se incrementan las probabilidades de estornudo. Un ejemplo de probabilidades que funcionarían para este caso es:

$$\begin{aligned}P(\text{estornudo}/\text{resfriado}, \text{rinitis}) &= 0.99 \\P(\text{estornudo}/\text{resfriado}, \sim \text{rinitis}) &= 0.85 \\P(\text{estornudo}/\sim \text{resfriado}, \text{rinitis}) &= 0.9 \\P(\text{estornudo}/\sim \text{resfriado}, \sim \text{rinitis}) &= 0.01\end{aligned}$$

Como se puede observar, las primeras tres probabilidades están cercanas a 1, mientras que la última está cercana a 0. Este modelo simple permite expresar varias cosas diferentes: la primera es que cuando hay más de una causa presente es más probable que el efecto ocurra (0.99 comparado a 0.85 y 0.9); la segunda es que la relación entre algunas causas es más fuerte que entre otras (por ejemplo, en este caso las probabilidades indican que la relación entre *rinitis* y *estornudo* es de alguna manera más fuerte que la relación entre *resfriado* y *estornudo*); la tercera es que, incluso cuando las dos causas están presentes, algo extraño puede ocurrir y la persona puede no estar estornudando (esta es la razón por la que la probabilidad es 0.99 y no 1). Los números 1 y 0 también se podrían usar para definir las probabilidades, indicando de este modo que no queremos modelar la incertidumbre presente en las relaciones.

El uso de modelos canónicos puede simplificar la especificación de parámetros. El lector interesado puede encontrar más información al respecto en (Diez, 2001).

Si procedemos de la misma forma con el resto de nodos podemos terminar el proceso de modelado. La figura siguiente muestra la RB completa, con nodos, enlaces y probabilidades. Nótese que en el caso de la *rinitis*, las probabilidades condicionales han sido definidas para modelar una relación tipo AND (la *rinitis* ocurre sólo cuando el gato y la alergia están presentes a la vez, excepciones aparte). En lo que sigue, para indicar las relaciones AND y OR, incluiremos una etiqueta en el modelo gráfico para dejar claro el tipo de relación existente entre los nodos. La red completa para el ejemplo del estornudo, con nodos, enlaces y parámetros se muestra en la Figura 8.

Cuando hay dos causas que por separado e independientemente entre ellas pueden causar el mismo síntoma, tenemos una relación tipo OR

Cuando hay dos causas que necesitan actuar conjuntamente para provocar cierto síntoma, tenemos una relación tipo AND

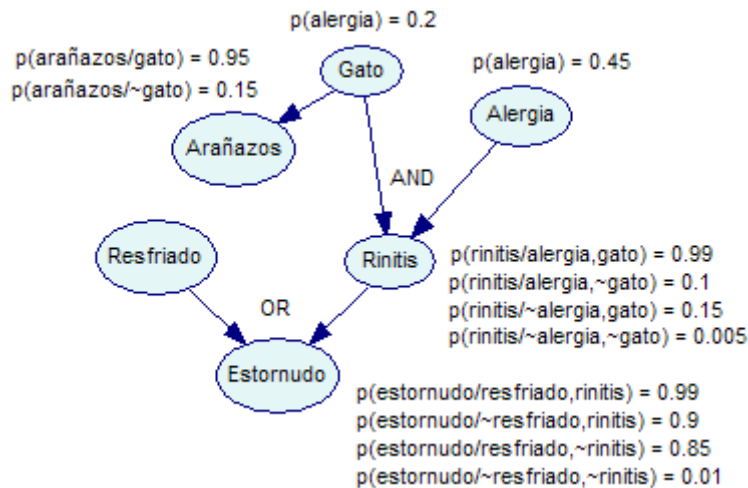


Figura 8. RB y parámetros para el ejemplo del resfriado.

3.5 Estructuras equivalentes

Como se explicó antes, desarrollar el modelo según las relaciones de causalidad simplifica el trabajo tanto en la deducción de la estructura como de los parámetros.

Cualquier enlace de la RB puede siempre ser invertido para desarrollar un modelo equivalente. Sin embargo, en el proceso de inversión se añaden nuevos enlaces. La figura siguiente muestra la red bayesiana para el ejemplo del estornudo, en la que se han invertido todos los enlaces, y los nuevos enlaces que ha sido necesario crear para que el comportamiento de la red sea equivalente. Asimismo, muestra a modo de ejemplo las probabilidades que se obtienen para uno de los nodos (el nodo gato). Como se puede observar, el modelo es mucho más complicado de interpretar y las probabilidades condicionadas más difíciles de estimar.

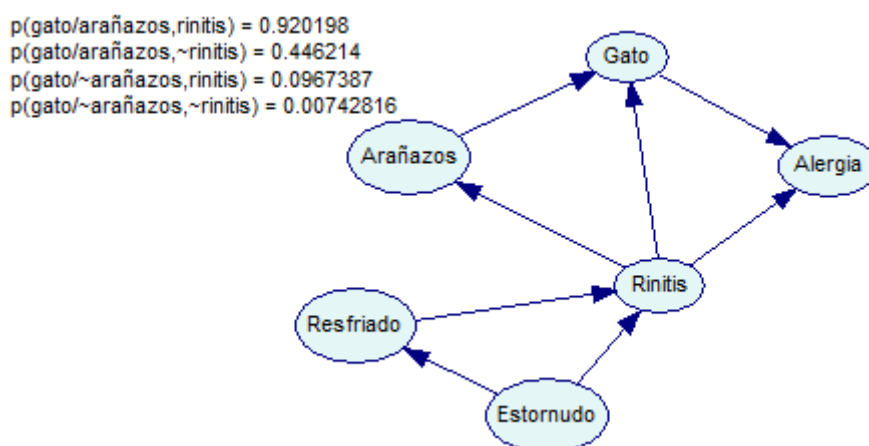


Figura 9. RB equivalente para el ejemplo del estornudo, con los enlaces invertidos

El conjunto de reglas de diagnóstico que explica este modelo es el siguiente:

- SI rinitis, ENTONCES arañazos
- SI rinitis, ENTONCES gato
- SI rinitis, ENTONCES alergia
- SI arañazos, ENTONCES gato
- SI refriado, ENTONCES rinitis
- SI estornudos, ENTONCES resfriado
- SI estornudos, ENTONCES rinitis
- SI gato, ENTONCES alergia

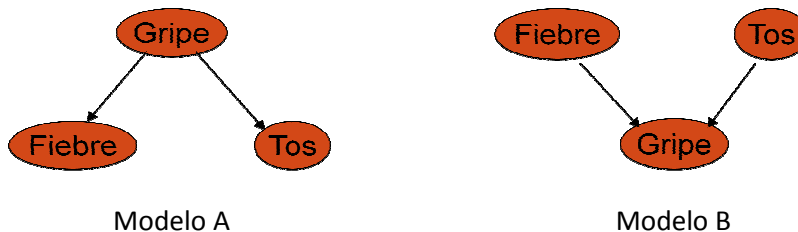
Como se puede ver, tanto el conjunto de reglas como el de probabilidades son de interpretación más compleja que el modelo causal.

3.6 Algunos trucos para el modelado

Hay una serie de trucos que en la práctica pueden resultar de utilidad. Mencionaremos algunos de ellos.

3.6.1 Verificar las independencias que supone el modelo

Una vez definido el modelo, para asegurarnos de que es correcto debemos comprobar si las relaciones entre las variables reflejan adecuadamente las dependencias e independencias existentes. Por ejemplo, para definir la relación entre las variables gripe, tos y fiebre tenemos dos posibles modelos:



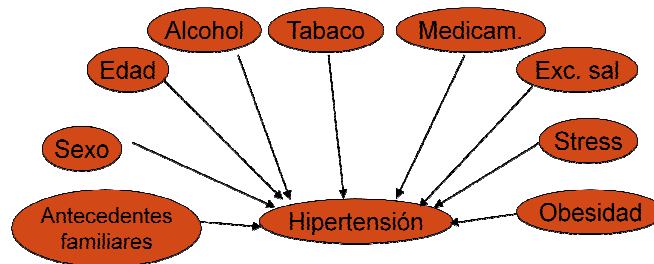
- En el modelo A, Fiebre y Tos son dependientes a priori pero independientes dado gripe
- En el modelo B, Fiebre y Tos son independientes a priori pero dependientes dado gripe (explaining-away).
- Vemos que en este caso el modelo que mejor refleja las independencias/dependencias que ocurren en la vida real es el A.

3.6.2 Introducir nodos intermedios para reducir la complejidad

Sabemos ya que los parámetros necesarios a la hora de definir una red son las probabilidades condicionadas de cada nodo dados sus padres. Por ello, el número de probabilidades necesarias para cada nodo es exponencial en el número de padres. De esta forma, cuando un nodo tiene muchos padres una forma de reducir la complejidad del modelo es introducir nodos intermedios que agrupen a varios padres. De este modo, no sólo reduciremos el número de

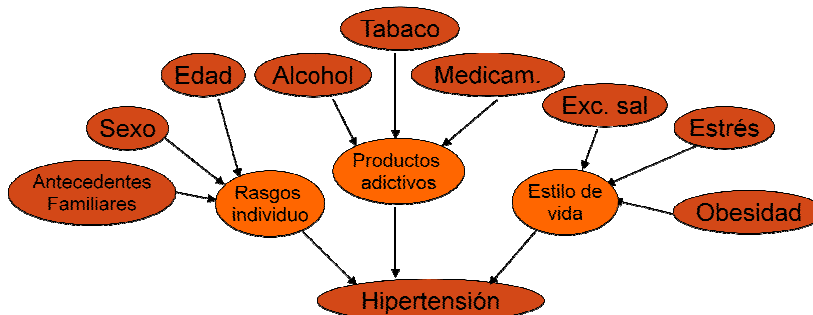
parámetros necesarios, sino también el tiempo de ejecución de los algoritmos de propagación de probabilidades.

Veamos un ejemplo. Consideremos el siguiente ejemplo: Supongamos que estamos construyendo un modelo para la hipertensión arterial, de la que un estudio ha identificado las siguientes causas: sexo, edad, antecedentes familiares, consumo de alcohol, tabaco, o ciertos medicamentos) excesiva ingesta de sal, obesidad, estrés. La red bayesiana para modelar esta situación sería:



Como vemos, en este modelo la hipertensión arterial tiene nueve padres, de modo que, en el caso de que los nodos sean binarios, para dar la probabilidad condicionada de la hipertensión necesitamos 2^9 valores.

Para simplificar el modelo, podemos intentar agrupar las causas del siguiente modo: edad, sexo y antecedentes familiares se pueden considerar rasgos propios del individuo; alcohol, tabaco y consumo de medicamentos se pueden agrupar en consumo de productos adictivos, mientras que exceso de sal en la dieta, estrés y obesidad pueden considerarse hábitos de vida. De este modo, si consideramos razonable este agrupamiento, el modelo sería:



En la que ahora para el nodo hipertensión necesitamos 23 parámetros, igual que para los nodos rasgos individuo, productos adictivos y estilo de vida. Es decir, que mediante esta técnica, hemos reducido el número de parámetros necesarios de 512 a 32.

3.6.3 Uso de modelos canónicos

Hemos visto que, en el caso general, el número de parámetros requerido para especificar la probabilidad condicional de un nodo crece exponencialmente con el número de padres, lo que plantea problemas tanto de obtención y almacenamiento de datos como de tiempos de computación de los algoritmos.

(por ejemplo, para un nodo binario con 10 padres binarios, la tabla de probabilidad condicional tiene $2^{10} = 2.048$ parámetros).

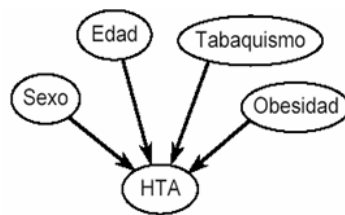
Por ello, es conveniente buscar modelos simplificados de interacción entre variables. Pearl los llama modelos *canónicos* por su aplicabilidad a diferentes dominios. Los más utilizados son el modelo NOISY-OR y NOISY-ADD.

En el caso de la puerta OR, es necesario realizar las siguientes hipótesis:

1. Cada una de las causas, por sí misma, puede producir el efecto
2. Basta que una de las causas produzca el efecto para que éste esté presente
3. Cuando todas las causas están ausentes, el efecto está ausente
4. No hay interacción entre las causas.

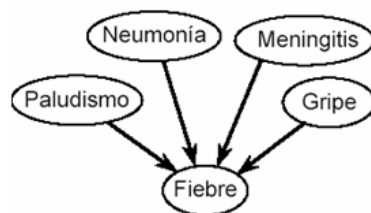
Para ilustrar situaciones en las que estas hipótesis son razonables y otras en las que no lo son, pensemos en los siguientes ejemplos:

Consideremos los siguientes factores, que tienen influencia causal en que una persona desarrolle hipertensión:



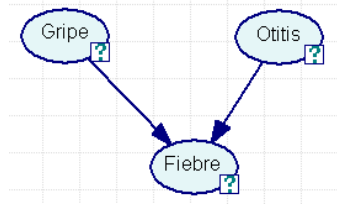
En este caso no podríamos utilizar una puerta OR, pues no se cumplen las hipótesis.

Consideremos los siguientes factores, y pensemos que son las únicas causas directas del síntoma Fiebre:



En este caso sí podríamos utilizar la puerta OR, pues se cumplen las hipótesis.

Estudiaremos este modelo de interacción causal a través de un ejemplo sencillo. Supongamos la siguiente red bayesiana:



Bajo estas hipótesis, para construir las probabilidades necesarias para el modelo bastaría con dar las probabilidades de que cada una de las causas provoque el efecto por separado (que denotaremos por c_x). Sea por ejemplo:

$$c_g = P(+f/+g) = 0.8$$

$$c_o = P(+f/+o) = 0.6$$

Y en ese caso tendríamos que:

$$P(+f/+g, +o) = 0.8 + 0.2 * 0.6 = 0.92$$

$$P(+f/-g, +o) = 0.6$$

$$P(+f/+g, -o) = 0.8$$

$$P(+f/-g, -o) = 0$$

Las probabilidades de que no se manifieste el síntoma pueden calcularse como complementarias de éstas, o aplicando la siguiente expresión:

$$P(\neg x/c_1, c_2) = \prod_{i \in T_u} (1 - c_i)$$

Donde T_u representa el conjunto de causas de U que están presentes. Así, tenemos que:

$$P(\neg f/+g, +o) = 0.2 * 0.4 = 0.08$$

$$P(\neg f/-g, +o) = 0.4$$

$$P(\neg f/+g, -o) = 0.2$$

$$P(\neg f/-g, -o) = 1$$

Supongamos que en el modelo queremos incluir que es posible que otras causas no determinadas provoquen también fiebre, introduciendo así cierto ruido en el mismo. Bajo las hipótesis anteriormente mencionadas, es posible construir las probabilidades condicionadas necesarias en el modelo a partir de unos parámetros básicos, concretamente las probabilidades de que cada causa provoque el efecto por separado y un factor de ruido r , que expresa que otras condiciones no presentes en el modelo podrían provocar la fiebre. Para continuar con el ejemplo, demos unos valores a estos parámetros:

$$c_g = P(+f/+g, -o, -r) = 0.8$$

$$c_o = P(+f/-g, +o, -r) = 0.6$$

$$r = p(+f/-g, -o) = 0.01$$

Veamos cómo se calculan en este caso las probabilidades condicionadas necesarias. En primer lugar, tenemos que por el axioma 4:

$$P(+f/+g, +o, -r) = P(+f/+g, -o, -r) + P(\neg f/+g, -o, -r) P(+f/-g, +o, -r) = 0.8 + 0.2 * 0.6 = 0.92$$

Por el axioma 3:

$$P(+f/-g, -o, -r) = 0$$

Las otras probabilidades se calculan mediante:

$$P(+f/+g, +o) = P(+f/+g, +o, -r) + P(-f/+g, +o, -r) P(+f/-g, -o) = 0,92 + 0,08*0.01 = 0,9208$$

$$P(+f/+g, -o) = P(+f/+g, -o, -r) + P(-f/+g, -o, -r) P(+f/-g, -o) = 0,8 + 0.2*0.01 = 0,802$$

$$P(+f/-g, +o) = P(+f/-g, +o, -r) + P(-f/-g, +o, -r) P(+f/-g, -o) = 0,6 + 0.4*0.01 = 0,604$$

En general, si en un modelo tenemos un efecto X y U_1, \dots, U_n son las causas posibles, si denotamos por c_i a la probabilidad que tiene cada causa de producir el efecto por separado y por q_i a la probabilidad complementaria ($q_i = 1 - c_i$), entonces:

$$P(\neg x/u_1, \dots, u_n) = \prod_{i \in T_u} q_i$$

Como en el caso anterior, el modelo se puede generalizar al caso de que haya cierto ruido, y en ese caso la probabilidad de cada combinación de estados se calculara como la probabilidad de que aparezca el efecto dada la combinación que sea más la probabilidad de que no aparezca el efecto dada la combinación que sea, multiplicada por el ruido.

Existen otros modelos de interacción causal que también están implementados en GeNIe:

- Noisy MAX: La generalización de la puerta OR al caso en que los las variables son multivaluadas. En este caso se considera que el efecto aparece a partir de cierto valor de la variable causa, y se procede de igual modo.
- Noisy ADDER: Según la ley de DeMorgan, el opuesto del Noisy-OR.

4 Razonamiento con redes bayesianas

Una vez definido formalmente el concepto de red bayesiana e introducido el proceso de modelado, pasamos a explicar cómo tiene lugar el proceso de razonamiento utilizando redes bayesianas. Para ello sólo vamos a ver el caso más simple, que es el algoritmo de propagación de probabilidades para el caso de redes con forma de árbol. Este caso nos servirá de ejemplo ilustrativo. Para el caso general y algoritmos aproximados, disponemos del software GeNIe.

4.1 Algoritmo de propagación de probabilidades para redes con forma de árbol³

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

Fase de inicialización

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

Fase de actualización

Cuando una variable se instancia, se actualiza el estado de la red, obteniéndose las probabilidades a posteriori de las variables de la red basadas en la evidencia considerada, adoptando la red un estado que denotaremos por S_1 .

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

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

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

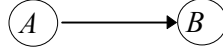
- La variable envía a su padre un mensaje, que llamaremos el λ -mensaje, para informarle de que ha cambiado su valor/probabilidad.
- La variable envía a todos sus hijos un mensaje, que llamaremos el π -mensaje, para informarlos de que ha cambiado su valor/probabilidad.

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

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

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

³ Tomado de [Neapolitan, 1990]



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

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

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

$$\begin{aligned}\lambda(A) &= (\lambda(a_1), \lambda(a_2), \lambda(a_3)) \\ \pi(A) &= (\pi(a_1), \pi(a_2), \pi(a_3))\end{aligned}$$

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

$$\begin{aligned}\lambda(B) &= (\lambda(b_1), \lambda(b_2)) \\ \pi(B) &= (\pi(b_1), \pi(b_2))\end{aligned}$$

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

Pasamos entonces a describir el algoritmo.

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

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

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

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

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

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

(*) Esta fórmula es válida en todos los casos. Otra fórmula de aplicación más sencilla, pero sólo es válida cuando todas las probabilidades $P^*(a_i)$ son no nulas, es $P(a_j) / \lambda_B(a_j)$. Proporciona un π -mensaje distinto (pero proporcional al de la otra fórmula) e iguales probabilidades a posteriori.

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

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

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

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

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

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

ALGORITMO:

1. Inicialización

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

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

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

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

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

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

2. Actualización

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

BEGIN

A.1. Inicializar $P^*(b_j) = 1$ y $P^*(b_i) = 0$, para todo $i \neq j$.

A.2. Calcular $\lambda(B)$ usando la fórmula 3.

A.3. Enviar un nuevo λ -mensaje al padre de B usando la fórmula 1.

A.4. Enviar nuevos π -mensajes a los hijos de B usando la fórmula 2.

END

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

BEGIN

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

END.

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

BEGIN

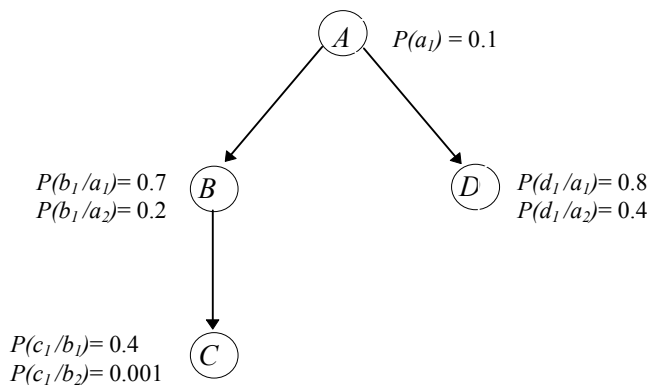
- C.1.** Calcular el nuevo valor de $\pi(B)$ usando la fórmula 4.
- C.2.** Calcular el nuevo valor de $P^*(B)$ usando la fórmula 5.
- C.3.** Enviar nuevos π -mensajes a los hijos de B usando fórmula 2.

END.

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

Ejemplo 7.

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



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

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

Inicialización

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

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

C. A envía un mensaje a su hijo, B,

$$\pi_B(a_1) = \pi(a_1)\lambda_D(a_1) = 0.1$$

$$\pi_B(a_2) = \pi(a_2)\lambda_D(a_2) = 0.9$$

B toma entonces nuevos π -valores;

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

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

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

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

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

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

$$\pi_C(b_1) = \pi(b_1) = 0.25$$

$$\pi_C(b_2) = \pi(b_2) = 0.75$$

Y actualiza su π -valor:

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

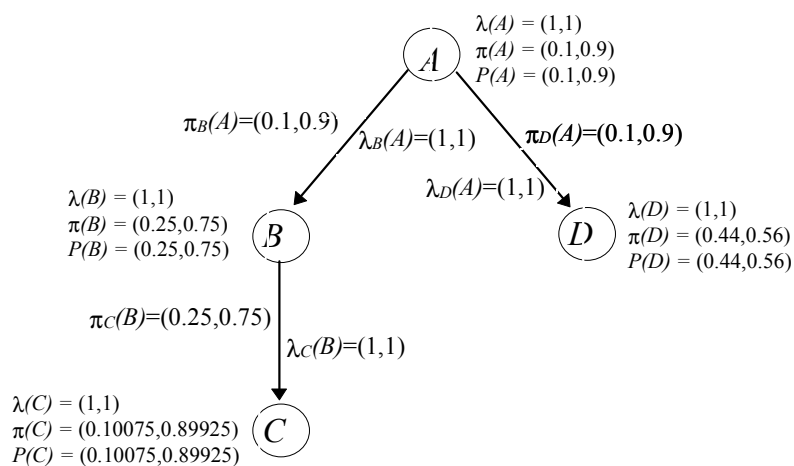
$$\pi(c_2) = P(c_2/b_1) \pi_C(b_1) + P(c_2/b_2) \pi_C(b_2) = 0.89925.$$

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

$$P(c_1) = 0.10075.$$

$$P(c_2) = 0.89925.$$

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



Estado S_0 de la red.

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

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

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

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

Veamos entonces cómo se efectúa la actualización con el algoritmo;

Actualización:

Actualización de B:

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

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

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

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

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

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

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

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

$$\lambda_B(a_2) = 0.2$$

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

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

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

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

Actualización de C :

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

C.1. El π -valor de C cambia,

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

$$\pi(c_2) = 0.6$$

C.2. Calculamos la nueva probabilidad de C

$$P^*(c_1) = 0.4 \propto 0.4$$

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

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

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

B.1. Actualizamos el λ -valor

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

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

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

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

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

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

B.3. A no tiene padre.

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

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

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

Actualización de D:

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

C.1. El π -valor de D cambia,

$$\pi(d_1) = 0.128$$

$$\pi(d_2) = 0.122$$

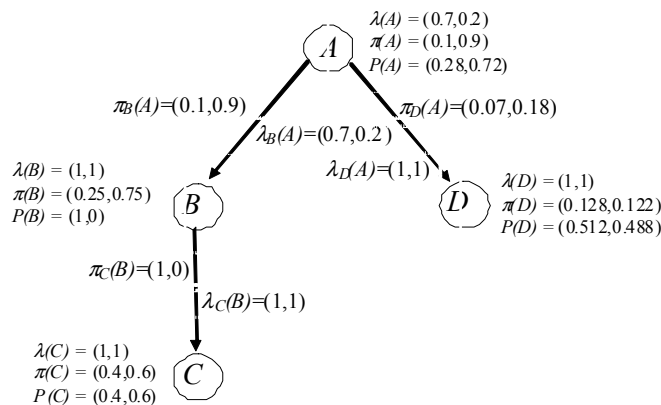
C.2. Calculamos la nueva probabilidad de D

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

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

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

Así, tras la instanciación de B a b1, la red queda;



Estado S_1 de la red.

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

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

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

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

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

Actualización:

Actualización de D

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

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

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

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

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

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

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

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

$$\lambda_D(a_2) = 0.6$$

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

Actualización de A

B.1. Calculamos $\lambda(A)$

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

$$\lambda(a_2) = \lambda_B(a_1) \lambda_C(a_2) = 0.2 \cdot 0.6 = 0.12$$

B.2. Calculamos la probabilidad actualizada de A

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

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

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

B.3. A no tiene padre.

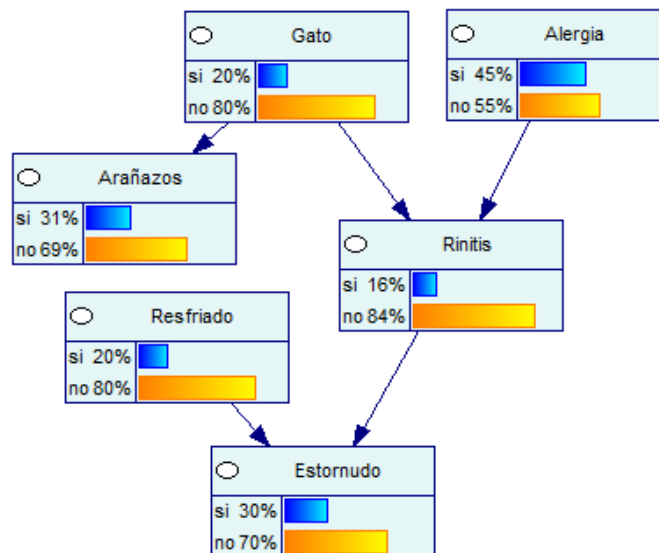
B.4. Este paso no se realiza pues B está ya instanciado.

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

4.2 Ejemplo de funcionamiento del caso general

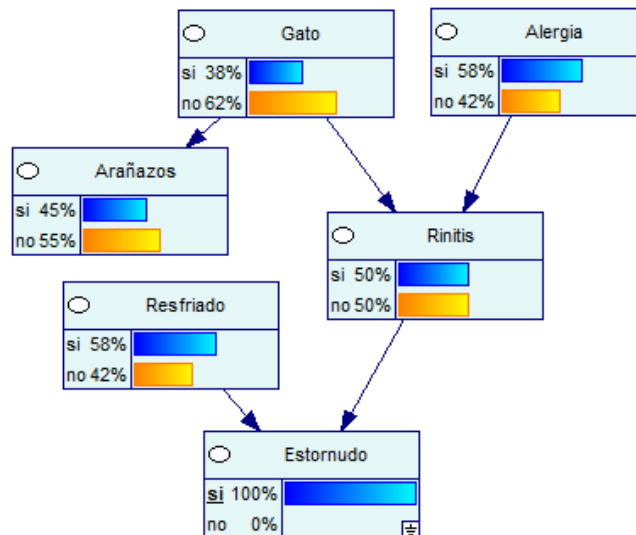
Como ya hemos mencionado, existen muchos algoritmos diferentes para el caso general, tanto exactos como aproximados, que quedan fuera del alcance de este curso. Pero para ilustrar su funcionamiento vamos a ver un ejemplo de razonamiento en el caso general, utilizando el software GeNIe. Como modelo utilizaremos el desarrollado anteriormente en el ejemplo del estornudo.

El primer paso es realizar el *proceso de inicialización* en el que se calculan todas las probabilidades a priori de todos los nodos de la red



Estado S_0 de la red para el ejemplo del estornudo

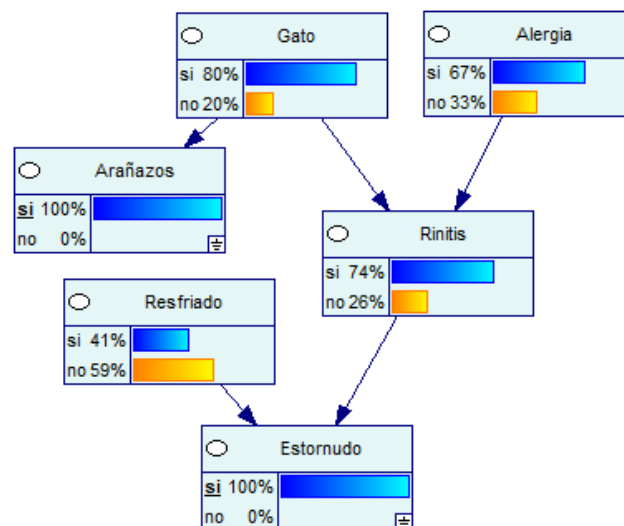
Cuando Juan empieza a estornudar, las probabilidades se actualizan para contabilizar esta información.



Estado S_1 de la red, tras conocer que Juan estornuda

Nótese que ahora las probabilidades de los estados positivos de todos los nodos de la red son mayores. Esto es así porque la evidencia disponible “Juan está estornudando” apoya el estado positivo de todas las variables. Recuerde que para el nodo estornudo, la única relación de independencia era que estornudo es independiente de gato, alergia y arañazos dado rinitis y resfriado (a priori, todos los nodos son dependientes de estornudo, y por eso cambian todas las probabilidades).

La siguiente evidencia disponible es que “Juan ve arañazos”. Las probabilidades actualizadas se pueden observar en la siguiente figura:



Estado S_2 de la red del ejemplo del estornudo.

Como puede observarse, ahora la probabilidad de gato se ha incrementado porque la evidencia arañazos favorece la presencia de gato. Al aumentar la probabilidad de gato, aumentan también las probabilidades de rinitis y alergia. La probabilidad de resfriado disminuye porque el estornudo puede explicarse ahora por la rinitis (efecto explaining-away).