Tema 2. Redes causales y redes bayesianas

UGR

Granada, noviembre/diciembre 2024

Índice

Imaginemos el siguiente problema:

- una mañana intento arrancar el coche y no lo consigo
- oigo cómo gira el motor de arranque, pero no hay forma
- puede haber varias causas para este problema. Ya que el motor de arranque funciona, debe haber batería. Por tanto, las causas más probables (en orden creciente de creencia) son:
  - que no haya combustible
  - que las bujías estén sucias
  - problema en el carburador
  - problema en sistema eléctrico
- para solucionar el problema compruebo el nivel de combustible. Al indicar que el depósito está medio, decido limpiar las bujías.

1. *Razonamiento en incertidumbre*

Para que un ordenador pueda imitar esta forma de razonar es necesario:

- conocer qué nos hace concluir que las causas más probables de fallo son la falta de combustible o la suciedad de las bujías (**conocimiento del problema**)
- qué hace que decida mirar el nivel de combustible (**conocimiento del problema**)
- cómo la observación del nivel de combustible cambia mi creencia en el estado de las bujías (**razonamiento**)

Para ser más precisos, es necesario:

- disponer de **conocimiento sobre el problema**
- formas de **representar el conocimiento**
- formas de **realizar inferencia** con la representación usada, de forma que el ordenador pueda simular el proceso de razonamiento

Una alternativa: lógica proposicional

- la forma de representación es la lógica booleana
- formas de razonamiento: tablas de verdad, diagramas de decisión, etc

En el razonamiento lógico se usan 4 formas de conectar hechos:

- conjunción: tanto Juan como María tienen gripe
- disyunción: o bien se quedan en casa o bien van al cine
- implicación: si llueve, el suelo está mojado
- negación: el suelo no está mojado

Estas son las reglas usadas para poder deducir (inferir) nuevo conocimiento a partir del ya disponible. Por ejemplo, de la tercera y cuarta sentencias podemos inferir que **no llueve**

1. *Razonamiento en incertidumbre*

Sin embargo, en la vida real hay incertidumbre... Para dotar a la lógica proposicional de la capacidad de tratamiento de la incertidumbre es necesario:

- dotar de nivel de certeza a las sentencias lógicas
- contar con mecanismos de inferencia que combinen de forma coherente factores de certeza o creencia:
  - *si* a *entonces* b *con certeza* 0.4
  - *si* b *entonces* c *con nivel de certeza* 0.6
  - dado que conozco la ocurrencia de a, ¿cuál es el valor de certeza en la ocurrencia de c?

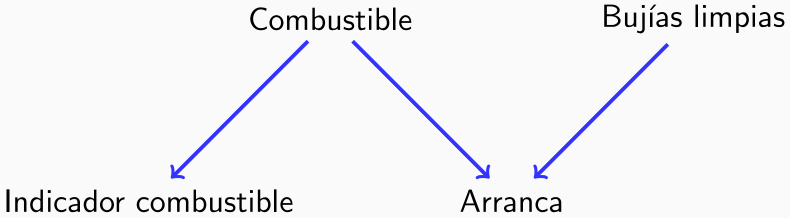
Otro problema adicional tiene que ver con una forma de razonamiento denominada **abducción**: si dispongo de la regla

> *una mujer tiene el pelo largo con nivel de certeza* 0.7

si veo una persona con pelo largo, ¿qué puedo inferir sobre su sexo? (es decir, se trata de una forma de razonamiento desde la observación hacia las causas)

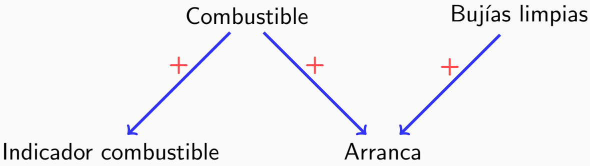
Otra alternativa: modelo gráfico. Se basa en construir un grafo que represente las relaciones causales entre los eventos identificados. Para simplificar el problema del arranque del coche, representamos la situación de la siguiente forma:

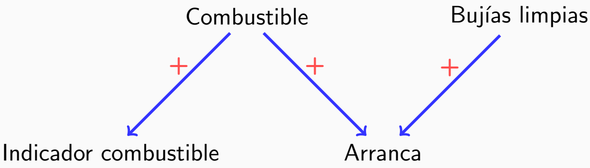
- **Combustible** toma dos posibles valores: sí, no
- **Bujías limpias** toma dos posibles valores: sí, no
- **Indicador combustible** toma tres posibles valores: lleno, medio, vacío
- **Arranca** tiene dos posibles valores: sí, no
- sabemos que conocer el valor de **Combustible** y **Limpiar bujías** tiene influencia causal en **Arranca**
- el estado de **Combustible** tiene impacto en el valor de **Indicador combustible**


Combustible
Bujías limpias
Indicador combustible
Arranca

Podemos decir además que las relaciones son positivas:

- a más certeza en que hay **Combustible**, más certeza en que el valor de **Indicador combustible** será mayor
- a más certeza de haber **Combustible** y de limpieza de las bujías, mayor certeza en que el coche arrancará





Este grafo puede usarse para razonar:

- si sé que las bujías no están limpias, aumenta mi certeza sobre problema de arranque (**razonamiento causal**, de causas a efectos)
- si sé que el coche no arranca, también aumenta mi creencia en las causas (**abducción** o **diagnóstico**, de efectos a causas)
- si el indicador marca la mitad de llenado, disminuye mi creencia en que el combustible sea la causa, lo que a su vez aumenta la creencia en problema de bujías (**razonamiento intercausal**, entre causas)

Índice

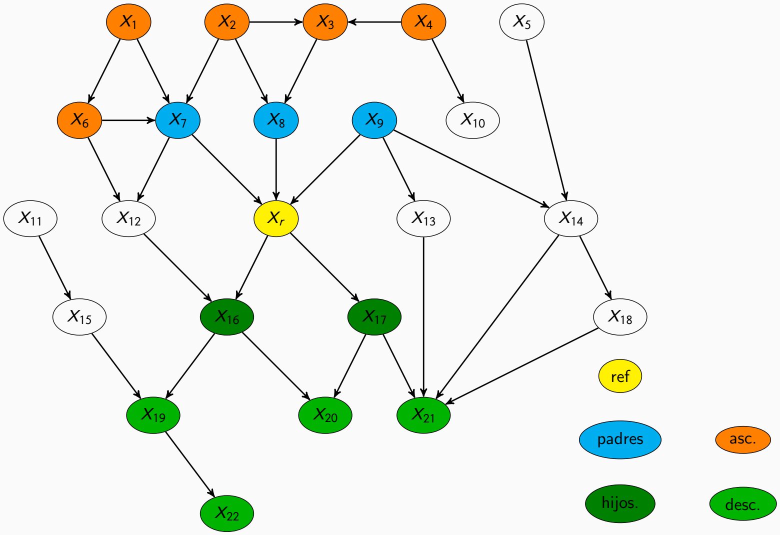
Visto el ejemplo anterior, podemos definir ahora qué es una red causal.

Red causal

una red causal consta de un conjunto de variables y un conjunto de enlaces dirigidos entre variables. Esta estructura se conoce como **grafo dirigido**. Si existe un enlace de A hasta B se dice que A es padre de B y B es hijo de A (también se habla de ascendientes y descendientes en general)

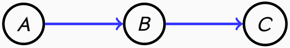
- las variables representan proposiciones o espacios muestrales
- una variable puede tener cualquier número de estados
- cada variable solo puede tomar a la vez uno de sus posibles valores

Ejemplo de red causal (red Bayesiana):

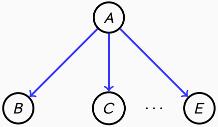


Es importante caracterizar los tipos de conexiones que pueden existir y la forma en que se transmite la información entre variables. Del tipo de conexión depende la forma de propagación de la información. Afortunadamente, solo puede haber tres tipos de conexiones:

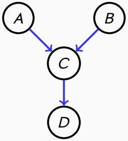
- serie



- divergente

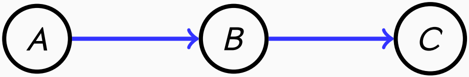


- convergente



Índice

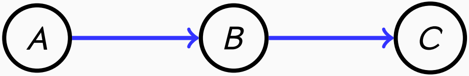
Consideremos la situación representada por la siguiente red causal:



- tener información sobre A afecta mi creencia sobre B y, a su vez, sobre C
- en sentido inverso, ocurre lo mismo
- ahora bien, si dispongo de información sobre B, entonces la evidencia sobre A ya no afecta a mi creencia sobre C

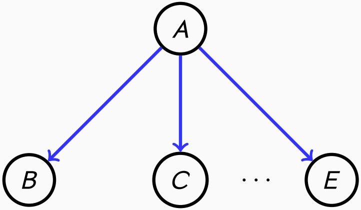
la información se transmite por las conexiones serie (en ambas direcciones) a menos que se conozca el valor de la variable B

Ejemplo concreto:



- A: lluvia
- B: nivel del agua
- C: inundación
- si no conozco el nivel del agua (B), saber que hubo inundación (C) cambia mi creencia en el nivel del agua y también en la posibilidad de que haya llovido (A) (conexión **abierta** y razonamiento abductivo)
- si conozco el nivel del agua, saber si llueve no aporta información para saber si habrá inundación o no (conexión **cerrada**)

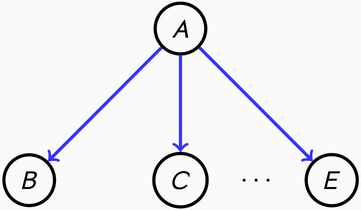
Índice



- si no conocemos el valor de A, tener información sobre el resto de variables modifica mi creencia sobre A
- si conozco el valor de A, entonces disponer de información sobre B (por ejemplo) no altera mi creencia sobre el resto de variables

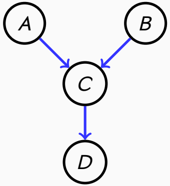
la información puede transmitirse en las conexiones divergentes a menos que se conozca el valor de A (que esté instanciada)

Ejemplo concreto:



- A (sexo); B (longitud del pelo); C (peso); E (altura)
- si no conocemos el sexo, tener información sobre la longitud de pelo modifica mi creencia sobre el sexo de la persona
- si sé que se trata de un hombre, entonces conocer la longitud del pelo no aporta información extra sobre peso o altura

Índice

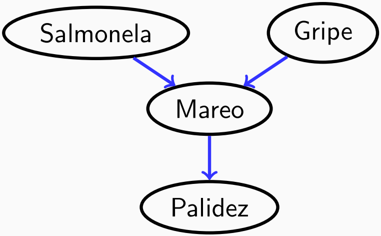


La situación es más complicada en las conexiones **convergentes**:

- si no dispongo de información sobre C o D entonces A y B son independientes entre sí. De esta forma, disponer de información sobre A no me aporta información adicional sobre B
- si tengo información sobre los valores de C o D, entonces disponer de información sobre A puede modificar mi creencia sobre B

en las conexiones convergentes hay flujo de información si y solo si hay evidencia sobre la variable en la conexión (v-estructura) o sobre alguna de sus sucesoras

Ejemplo concreto:



- si no tengo información sobre la presencia de mareo o palidez, entonces la información sobre salmonela no afecta mi creencia sobre gripe
- si sabemos que la persona está pálida, entonces saber que no tiene salmonela incrementa mi creencia en la ocurrencia de gripe
- si sabemos que la persona está pálida, saber que tiene salmonela disminuye mi creencia en la ocurrencia de gripe

Índice

Los tipos de conexiones vistos anteriormente son todos los que pueden aparecen en una red causal, en un grafo dirigido. Mediante el siguiente criterio podemos determinar si dos variables son independientes dada una cierta evidencia.

d-separación

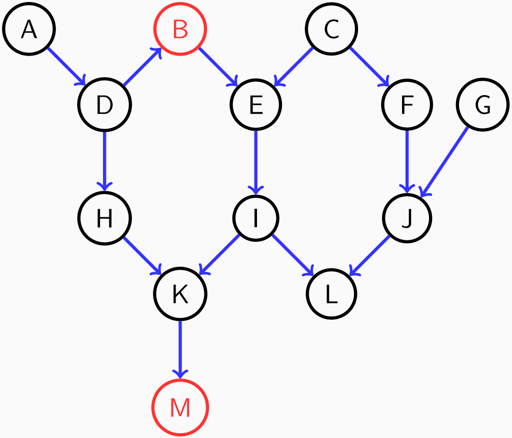
Dos variables A y B están **d-separadas** (d - grafo **d**irigido) en una red causal si para todos los posibles caminos entre ellas existe una variable intermedia V (distinta de A y B) tal que:

- la conexión es serie o divergente y V tiene valor conocido
- la conexión es convergente en V y no se dispone de información sobre V ni sobre ninguna de sus sucesoras

Si A y B no están **d-separadas** entonces están **d-conectadas** (d-separación implica independencia).

caminos cerrados (hay **d-separación**) / abiertos (**d-conexión**).

Ejemplo 1:

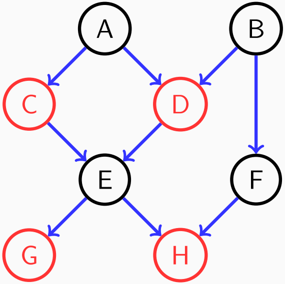


d-separación $A - L$?

d-separación $A - C$?

d-separación $A - G$?

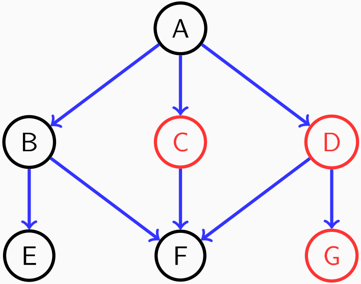
Ejemplo 2:



d-separación $A - F$?

d-separación $A - E$?

Ejemplo 3:



d-separación $A - F$?
d-separación $F - E$?
d-separación $F - G$?

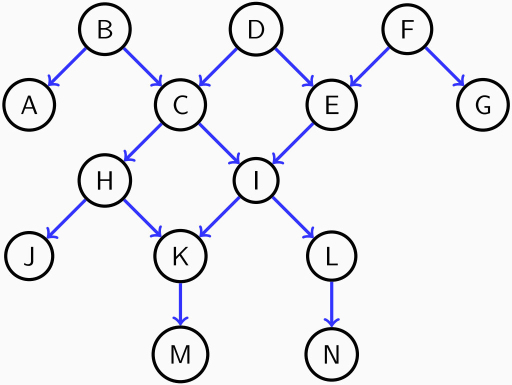
Índice

Manto de Markov

El manto de Markov de una variable A es el conjunto formado por:

- los padres de A
- los hijos de A
- aquellas variables que también tienen un hijo común con A

Si se instancian las variables del manto de Markov de A, entonces A está d-separada del resto de variables de la red.



¿manto de Markov de I?

Índice

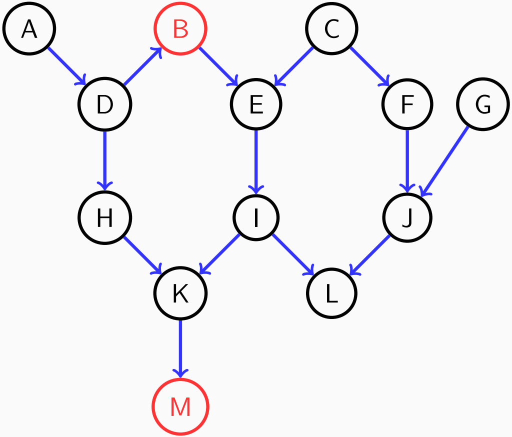
Grafo moral

otra forma más sencilla de determinar la d-separación entre A y B dada la evidencia de un conjunto de variables $\mathcal{S}$ consiste en:

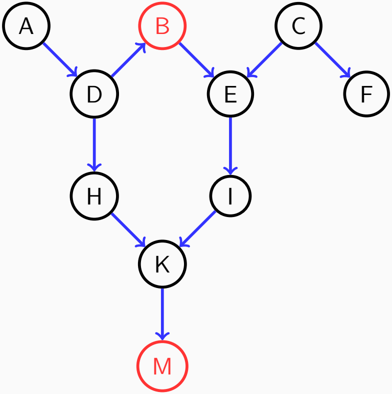
- construir el grafo ancestral formado por A, B y $\mathcal{S}$ (contiene todas estas variables y aquellas otras con caminos dirigidos a alguna de ellas)
- se inserta un enlace no dirigido entre cada par de nodos con un hijo en común y se eliminan las direcciones de los enlaces

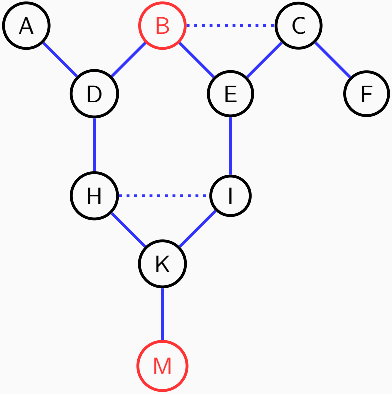
Este grafo se conoce como **grafo moral**.

Sobre él puede determinarse la **u-separación** (equivalente a d-separación) entre dos nodos A y B dada la evidencia $\mathcal{S}$; existe separación si todos los caminos entre A y B pasan por algún nodo en $\mathcal{S}$.



Se usa este procedimiento para este problema para determinar la d-separación entre A y F dada evidencia para $C = \{B, M\}$





Grafos ancestral y moral: se observa fácilmente que hay caminos que permiten conectar A y F.

Índice

Definición de red bayesiana

Una **red bayesiana** consta de los siguientes elementos:

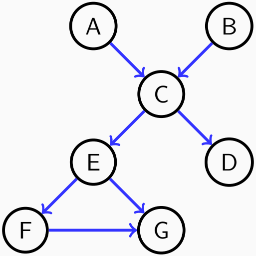
- un conjunto de variables y un conjunto de enlaces dirigidos entre variables
- cada variable tiene un conjunto finito de estados exclusivos
- las variables junto con los enlaces dirigidos forman un **grafo dirigido acíclico** (no hay en camino como $A_1 \rightarrow \ \ldots. \ \rightarrow A_n$, tal que $A_1 = A_n$)
- se asigna a cada variable A con padres B_1 .... B_n una tabla con los valores de la distribución condicional $P(A|B_1 \ldots. B_n)$

La definición no incluye la noción de causalidad y no se precisa que los enlaces representen relaciones **causa-efecto**

Lo importante es comprobar que se cumplen las condiciones de **d-separación**. Si A y B están d-separadas dada la evidencia e, debe cumplirse que:

$$P(A|e) = P(A|B, e)$$

Es decir, la información cualitativa (grafo) y cuantitativa (distribuciones de probabilidad) deben codificar las mismas relaciones de dependencia e independencia



Distribuciones asignadas:

- P(A)
- P(B)
- P(C|A, B)
- P(D|C)
- P(E|C)
- P(F|E)
- P(G|E, F)

En relación a la forma de hacer inferencia, supongamos un conjunto de variables $\mathcal{U} = \{X_1 \; .... \; X_n\}$. Si tenemos acceso a la distribución conjunta $P(\mathcal{U})$ entonces podamos realizar sobre ella todos los cálculos que necesitemos, con o sin evidencia.

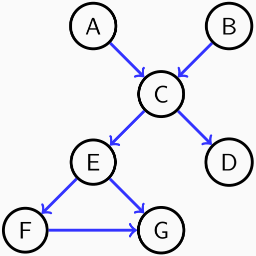
El problema radica en que el tamaño de la distribución (el número de parámetros que contiene) crece de forma exponencial con el número de variables. No hace falta un número muy grande de variables para llegar a tamaños que son manejables.

En definitiva, necesitamos representaciones compactas para $P(\mathcal{U})$. Esta ventaja se obtiene con las redes bayesianas (ya que evitan la necesidad de operar de forma directa con la distribución conjunta)

Sea R una red bayesiana definida sobre un conjunto de variables $\mathcal{U}$ y sea $P(\mathcal{U})$ la distribución conjunta definida por la red. Esta distribución conjunta puede obtenerse mediante la aplicación de la **regla de la cadena**:

$$P(\mathcal{U}) = \prod_{i=1}^{n} P(X_i | pa(X_i))$$

donde $pa(X_i)$ denota los padres de la variable X_i en la red. Es decir, podemos generar la conjunta a partir de las distribuciones individuales (y, en general, no será necesario el uso de la conjunta y bastará con operar con las distribuciones de las variables d-conectadas)

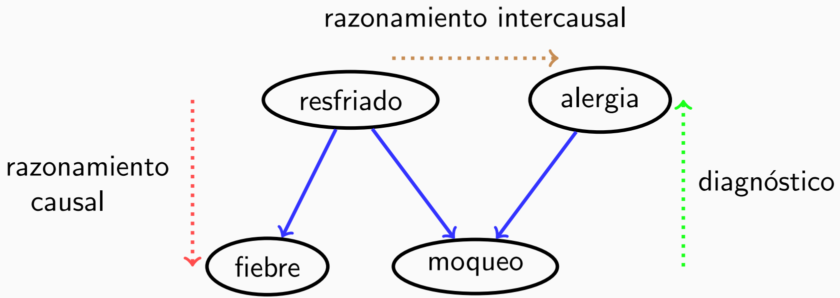


$$P(\mathcal{U}) = P(A)P(B)P(C|A,B)P(D|C)P(E|C)P(F|E)P(E|F,G)$$

Además, gracias a la **d-separación** sabemos que hay variables que pueden no influir en algunos cálculos.

Ejercicio: ¿qué distribuciones serían necesarias para calcular $P(A,B|E)$?

Como ya se ha indicado antes, en las redes bayesianas se pueden llevar a cabo diferentes formas de razonamiento:



Índice

Mucha veces interesa conocer la forma en que disponer de información (**evidencia**) modifica mi creencia sobre la probabilidad de ocurrencia de la variables de la red. Se suele representar mediante $\boldsymbol{e}$ para indicar que es un vector, que puede contener el valor de varias variables.

Para poder calcular necesitamos eliminar de las distribuciones de probabilidad todas aquellas combinaciones que no son coherentes con la evidencia. Para trabajar de forma general asumimos que la evidencia sobre una variable X con n estados se representa con una tabla de ceros y unos. Los ceros se corresponden con los valores no observados. Llamaremos a este potencial $\phi_e(X)$.

Imaginemos la distribución $P(A|B, C)$, donde todas las variables tienen dos estados posibles. Podría ser la siguiente:

BC	b_1c_1	b_1c_2	b_2c_1	b_2c_2
$A = a_1$	0.3	0.6	0.9	0.8
$A = a_2$	0.7	0.4	0.1	0.2

y queremos incorporar la evidencia $A = a_1$. Para incorporar la evidencia se usa un potencial para representarla, que llamaremos $\phi_{A=a_1}(A, B, C)$:

BC	b_1c_1	b_1c_2	b_2c_1	b_2c_2
$A = a_1$	1	1	1	1
$A = a_2$	0	0	0	0

Se observa que solo hay valores 1 asociado a las configuraciones en que $A = a_1$.

De esta forma, incorporar la evidencia implica hacer la multiplicación:

$$P(A|B,C)\phi_{A=a_1}(A,B,C)$$

Y esta multiplicación produce como resultado un potencial que solo depende de B y C, que denominamos $\phi_2(B,C)$ y que se representa abajo:

BC	b_1c_1	b_1c_2	b_2c_1	b_2c_2
	0.3	0.6	0.9	0.8

Hablamos de potencial porque no se cumplen las propiedades de las distribuciones, es decir, los valores no suman 1.

Teniendo en cuenta la forma de incorporar la evidencia, si disponermos de una red bayesiana R definida sobre $\mathcal{U}$ y de la evidencia $\boldsymbol{e}$ entonces podemos calcular cómo varía la creencia sobre el resto de variables mediante el cálculo:

$$P(\mathcal{U}, \boldsymbol{e}) = P(\mathcal{U})\phi_{\boldsymbol{e}}$$

Otro cálculo de interés es el de la probabilidad de la evidencia:

$$P(e) = \sum_{\mathcal{U}} P(\mathcal{U}, \boldsymbol{e}) = \sum_{\mathcal{U}} P(\mathcal{U})\phi_{\boldsymbol{e}}$$

Sea R una red bayesiana definida sobre $\mathcal{U}$, sea $\boldsymbol{e}$ la evidencia disponible y sean los potenciales ϕ_{e_1} .... ϕ_{e_m} que definen las configuraciones observadas para las variables consideradas en $\boldsymbol{e}$. Entonces:

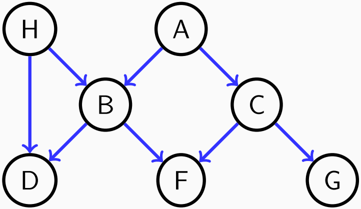
$$P(\mathcal{U}, \boldsymbol{e}) = \prod_{X \in \mathcal{U}} P(X|pa(X)) \prod_{i=1}^{m} \phi_{e_i}$$

Para cada variable X podemos calcular la distribución **a posteriori** a la luz de la evidencia disponible:

$$P(X|e) = \frac{\sum_{\mathcal{U} \setminus X} P(\mathcal{U}, e)}{P(e)}$$

Índice

Consideremos la red:



Cada variable tiene 10 posibles valores (denotados de a a j) y la evidencia sobre la red es $\boldsymbol{e} = \{D = d_i, F = f_j\}$. Mediante la regla de la cadena:

$$P(\mathcal{U}, e) = P(A)P(B|H, A)P(C|A)P(D|B, H)\phi_{D=d_i}(B, D, H)$$
$$P(F|B, C)\phi_{F=f_j}(F, B, C)P(G|C)P(H)$$

Se han remarcado en rojo los potenciales de 0's y 1's con la información de la evidencia.

El primer paso para cualquier cálculo consistiría en incorporar la evidencia, haciendo las multiplicaciones vistas previamente:

$$P(D|B,H)\phi_{D=d_i}(D,B,H) = \phi_1(B,H)$$
$$P(F|B,C)\phi_{F=f_j}(F,B,C) = \phi_2(B,C)$$

En definitiva, el conjunto de potenciales disponibles tras este paso es:

$$\Phi = \{P(A), P(B|H,A), P(C|A), \phi_1(B,H), \phi_2(B,C), P(G|C), P(H)\}$$

Si el cálculo se hiciese directamente sobre la distribución conjunta necesitaría manejar 10^7 valores (10 millones de valores.....). Gracias a la descomposición mostrada por la regla de la cadena, el número de valores que se necesitan es mucho menor y disminuye aún más, en caso de haber evidencia.

dist.	sin evid.	con evid.
$P(A)$	10	10
$PB\|H, A)$	1000	1000
$P(C\|A)$	100	100
$P(D\|B, H)$	1000	100
$P(F\|B, C)$	1000	100
$P(G\|C)$	100	100
$P(H)$	10	10
total	3220	1420

Imaginemos que deseamos calcular $P(A, \boldsymbol{e})$. Para ello hay que eliminar las variables B, C y H. El orden de realización de las operaciones no afecta al resultado obtenido (pero sí a la complejidad del proceso). Usaremos exactamente ese orden.

Se comienza marginalizando B. Se retiran de Φ los potenciales que contienen a la variable B en su dominio, para poder eliminarla. Los potenciales seleccionados se incorporan a un conjunto que llamamos $\mathcal{I}$ De esta forma:

$$\Phi = \{P(A), P(C|A), P(H)\}$$
$$\mathcal{I} = \{P(B|A, H), \phi_1(B, H), \phi_2(B, C)\}$$

Hay que multiplicar todos los potenciales en $\mathcal{I}$ y después marginalizar la variable a borrar. En este caso:

$$\sum_B P(B|A,H)\phi_1(B,H)\phi_2(B,C) = \phi_3(A,C,H)$$

El potencial resultante se incorpora a Φ.

Tras la operación de borrado de B, tenemos:

$$\Phi = \{P(A), P(C|A), \phi_3(A, C, H), P(H)\}$$

Ahora se marginaliza la variable C y se procede de la misma forma: se retiran de Φ los potenciales con C y estos se incorporan a $\mathcal{I}$:

$$\Phi = \{P(A), P(H)\}$$
$$\mathcal{I} = \{P(C, A), \phi_3(A, C, H)\}$$

Hay que multiplicar todas las distribuciones incluidas en $\mathcal{I}$:

$$P(C, A)\phi_3(A, C, H) = \phi_4(A, C, H)$$

Y marginalizar la variable C:

$$\sum_C \phi_4(A, C, H) = \phi_4(A, H)$$

De momento tenemos:

$$\Phi = \{P(A), \phi_4(A, H), P(H)\}$$

La siguiente operación consiste en eliminar H:

$$\Phi = \{P(A)\}$$
$$\mathcal{I} = \{\phi_4(A, H), P(H)\}$$

Como en la operación anterior, hay que multiplicar los potenciales en $\mathcal{I}$ para luego marginalizar H:

$$\phi_4(A, H)P(H) = \phi_5(A, H)$$
$$\sum_H \phi_5(A, H) = \phi_6(A)$$

Tras realizar todos los borrados queda un conjunto de potenciales definidos únicamente sobre A:

$$\Phi = \{P(A), \phi_6(A)\}$$

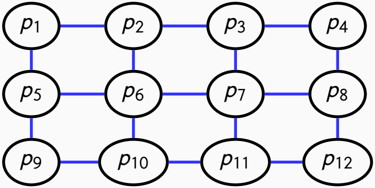
Se realiza su multiplicación $P(A)\phi_6(A) = \phi_R(A)$. Este es el potencial resultante buscado, que corresponde a $P(A, \boldsymbol{e})$. Este procedimiento se denomina **eliminación de variables** y, como hemos visto, evita el uso de la distribución conjunta.

Se observa que ninguna operación ha generado potenciales con más de tres variables.

Índice

Podemos tener modelos gráficos tanto **no dirigidos** como **dirigidos**. En los primeros, no se puede determinar la dirección de la relación (de ahí que se prescinda de la dirección de los arcos).

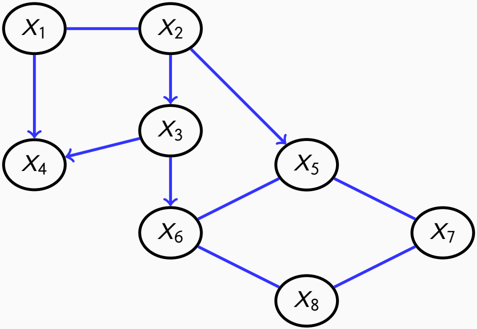
Por ejemplo, podemos pensar en un problema de análisis de imágenes. Está claro que existe relación entre pixels en posiciones vecinas, pero, ¿en qué sentido?



Las distribuciones asociadas al modelo serían:

$$P(\mathcal{U}) = \phi_1(p_1, p_2)\phi_2(p_1, p_5)\phi_3(p_2, p_3) \; .... \; \phi_n(p_{11}, p_{12})$$

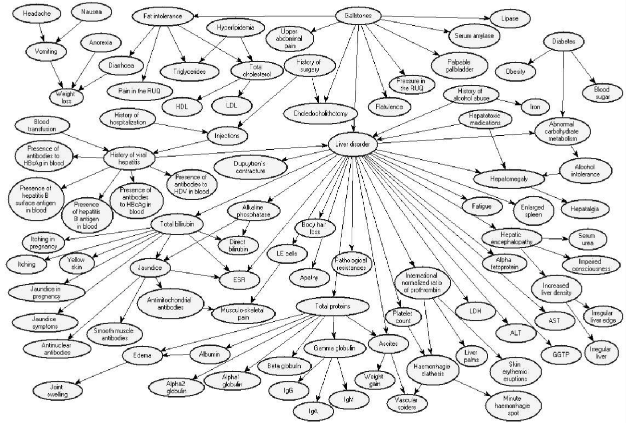
También hay modelos mixtos, donde solo algunos de los enlaces están dirigidos. Estos modelos se denominan **grafos cadena**.



El conjunto de distribuciones de este problema sería:

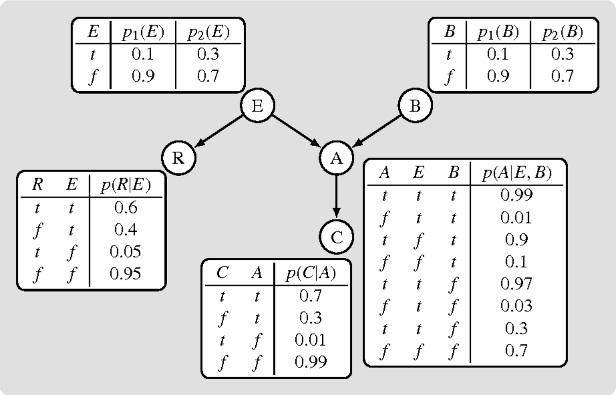
$$P(\mathcal{U}) = \phi_1(X_1, X_2)P(X_3|X_2)P(X_4|X_1, X_3)P(X_5|X_2)$$
$$P(X_6, X_5|X_3)\phi_2(X_5, X_6, X_7, X_8)$$

El modelo más conocido y usado son las redes bayesianas, representadas por grafos dirigidos acíclicos.



Ejemplo tomado de *A Bayesian Network Model for Diagnosis of Liver Disorders. A. Onisko, M.J. Druzdzel, H. Wasyfuk*

Las redes credales usan el mismo tipo de representación que las redes bayesianas, pero permiten imprecisión en la asignación de probabilidades



E	$p_1(E)$	$p_2(E)$
t	0.1	0.3
f	0.9	0.7

B	$p_1(B)$	$p_2(B)$
t	0.1	0.3
f	0.9	0.7

R	E	$p(R\|E)$
t	t	0.6
f	t	0.4
t	f	0.05
f	f	0.95

C	A	$p(C\|A)$
t	t	0.7
f	t	0.3
t	f	0.01
f	f	0.99

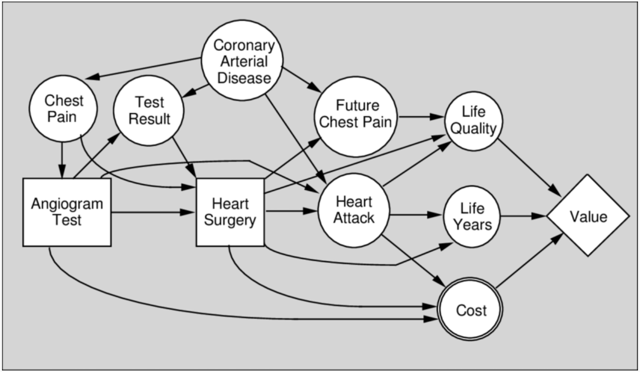
A	E	B	$p(A\|E,B)$
t	t	t	0.99
f	t	t	0.01
t	f	t	0.9
f	f	t	0.1
t	t	f	0.97
f	t	f	0.03
t	t	f	0.3
f	f	f	0.7

Ejemplo tomado de *Updating credal networks is approximable in polynomial time.*

D. Mauá, C.P. de Campos, M. Zaffalon

4. *Tipos de modelos gráficos*

Los diagramas de influencia incorporan más tipos de nodos y permiten representar y evaluar problemas de toma de decisiones



Ejemplo tomado de *Decision Analysis and Expert Systems. A. M. Henrion, J.S. Breese, E. Horvitz*

4. *Tipos de modelos gráficos*

Ventajas de los modelos gráficos probabilísticos:

- permiten la fácil comunicación con los expertos en el dominio del problema gracias a su representación gráfica
- el modelo es fácil de convertir en alguna estructura que pueda manejarse mediante programas de ordenador
- ofrecen la posibilidad de realizar diferentes tipos de razonamiento
- permiten el tratamiento de problemas complejos, con muchas variables, al no necesitar la distribución conjunta
- ofrecen un marco coherente para manipular la incertidumbre, gracias al uso de la teoría de la probabilidad