Teoría de probabilidades

- Definiremos como **probabilidad a priori** ($P(a)$) asociada a una proposición como el grado de creencia en ella a falta de otra información
- Definiremos como **probabilidad a posteriori** o **condicional** ($P(a|b)$) como el grado de creencia en una proposición tras la observación de proposiciones asociadas a ella
- La probabilidad a posteriori se puede definir a partir de probabilidades a priori como:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

- Esta fórmula se puede transformar en lo que denominaremos la **regla del producto**:

$$P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$$

Inferencia probabilística

Las leyes de la probabilidad permiten establecer diferentes métodos de inferencia

- **Marginalización:** Probabilidad de una proposición atómica con independencia de los valores del resto de proposiciones

$$P(Y) = \sum_z P(Y, z)$$

- **Probabilidades condicionadas:** Probabilidad de una proposición dados unos valores para algunas proposiciones e independiente del resto de proposiciones (a partir de la regla del producto)

$$P(X|e) = \alpha \sum_y P(X, e, y)$$

El valor α es un factor de normalización que corresponde a factores comunes que hacen que las probabilidades sumen 1

Inferencia probabilística: ejemplo

Consideremos un problema en el que intervengan las proposiciones
$Fumador = \{fumador, \neg fumador\}$, $Sexo = \{varon, mujer\}$,
$Enfisema = \{enfisema, \neg enfisema\}$

	enfisema		$\neg$*enfisema*	
	varon	*mujer*	*varon*	*mujer*
fumador	0.2	0.1	0.05	0.05
$\neg$*fumador*	0.02	0.02	0.23	0.33

Inferencia probabilística: ejemplo

$$
\begin{aligned}
P(\mathit{enfisema} \wedge \mathit{varon}) &= 0{,}2 + 0{,}02 \\
P(\mathit{fumador} \vee \mathit{mujer}) &= 0{,}2 + 0{,}1 + 0{,}05 + 0{,}05 + 0{,}02 + 0{,}33 \\
P(\mathit{Fumador}|\mathit{enfisema}) &= \langle P(\mathit{fumador}, \mathit{enfisema}, \mathit{varon}) \\
&\quad + P(\mathit{fumador}, \mathit{enfisema}, \mathit{mujer}), \\
&\quad P(\neg \mathit{fumador}, \mathit{enfisema}, \mathit{varon}) \\
&\quad + P(\neg \mathit{fumador}, \mathit{enfisema}, \mathit{mujer}) \rangle \\
&= \alpha \langle 0{,}3, 0{,}04 \rangle \\
&= \langle 0{,}88, 0{,}12 \rangle
\end{aligned}
$$

Inferencia probabilística: Problemas

- Hacer estos procesos de inferencia requiere almacenar y recorrer la distribución de probabilidad conjunta de todas las proposiciones
- Suponiendo proposiciones binarias el coste en espacio y tiempo es $O(2^n)$ siendo n el número de proposiciones
- Para cualquier problema real estas condiciones son impracticables
- Necesitamos mecanismos que nos simplifiquen el coste del razonamiento

Independencia probabilística

- Por lo general no todas las proposiciones que aparecen en un problema están relacionadas entre si
- Muestran la propiedad que denominaremos **independencia probabilística**
- Esto quiere decir que unas proposiciones no influyen en las otras y por lo tanto podemos reescribir sus probabilidades como:

$$P(X|Y) = P(X); \quad P(Y|X) = P(Y); \quad P(X,Y) = P(X)P(Y)$$

- Dadas estas propiedades podremos reescribir las probabilidades conjuntas de manera mas compacta reduciendo la complejidad

La regla de Bayes

- Hemos enunciado la regla del producto como:

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

- Esto nos lleva a lo que denominaremos la **regla de Bayes**

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- Esta regla y la propiedad de independencia serán el fundamento del razonamiento probabilístico y nos permitirá relacionar las probabilidades de unas evidencias con otras

La regla de Bayes + independencia

- Suponiendo que podemos estimar exhaustivamente todas las probabilidades que involucran la variable Y podemos reescribir la formula de Bayes como:

$$P(Y|X) = \alpha P(X|Y)P(Y)$$

- Suponiendo independencia condicional entre dos variables podremos escribir:

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

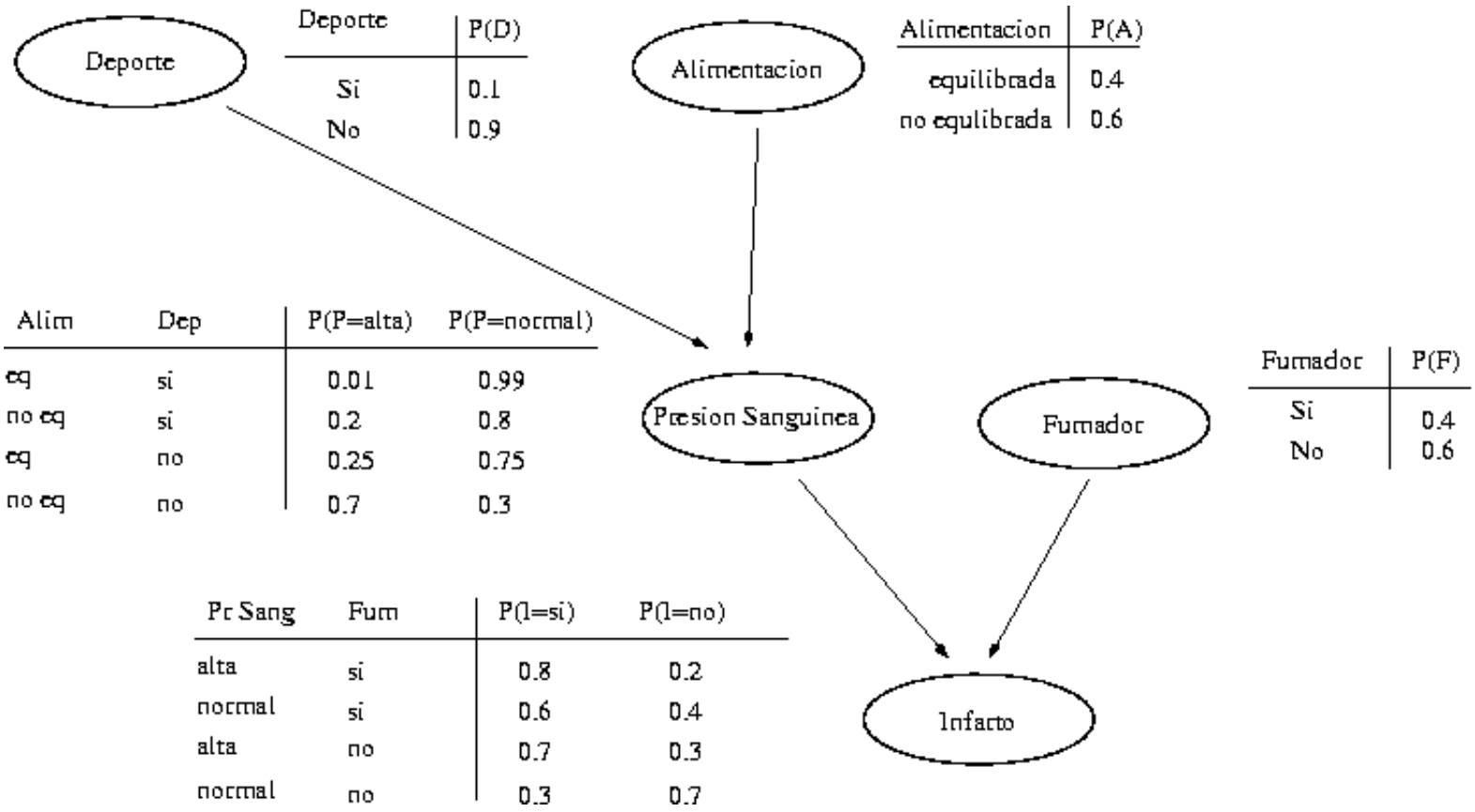
- De manera que:

$$P(Z|X, Y) = \alpha P(X|Z)P(Y|Z)P(Z)$$

Redes Bayesianas

- Si determinamos la independencia entre variables podemos simplificar el cálculo de la combinación de sus probabilidades y su representación
- Las **redes bayesianas** permiten la representación de las relaciones de independencia entre variable aleatorias
- Una red bayesiana es un **grafo dirigido acíclico** que tiene información probabilística en sus nodos indicando cual es la influencia de sus padres en el grafo sobre el nodo ($P(X_i|padres(X_i))$)
- El significado intuitivo de un enlace entre dos nodos X e Y es que la variable X tiene influencia sobre Y
- El conjunto de probabilidades representadas en la red describe la distribución de probabilidad conjunta de todas las variables

Redes Bayesianas: ejemplo



Deporte	P(D)
Si	0.1
No	0.9

Alimentacion	P(A)
equilibrada	0.4
no equlibrada	0.6

Alim	Dep	P(P=alta)	P(P=normal)
eq	si	0.01	0.99
no eq	si	0.2	0.8
eq	no	0.25	0.75
no eq	no	0.7	0.3

Fumador	P(F)
Si	0.4
No	0.6

Pr Sang	Fum	P(I=si)	P(I=no)
alta	si	0.8	0.2
normal	si	0.6	0.4
alta	no	0.7	0.3
normal	no	0.3	0.7

Redes Bayesianas - Distribución conjunta

- En cada nodo de la red aparece la distribución de probabilidad del nodo respecto a sus padres
- Esto permite factorizar la distribución de probabilidad conjunta, convirtiéndose en el producto de probabilidades condicionales independientes

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | padres(x_i))$$

Redes Bayesianas - Distribución conjunta - ejemplo

$$
\begin{aligned}
&P(Infarto = si \wedge Presion = alta \wedge Fumador = si \\
&\wedge Deporte = si \wedge Alimentacion = equil) \\
&= \\
&P(Infarto = si | Presion = alta, Fumador = si) \\
&P(Presion = alta | Deporte = si, Alimentacion = equil) \\
&P(Fumador = si) P(Deporte = si) P(Alimentacion = equil) \\
&= 0{,}8 \times 0{,}01 \times 0{,}4 \times 0{,}1 \times 0{,}4 \\
&= 0{,}000128
\end{aligned}
$$

Construcción de redes bayesianas

- Las propiedades de las redes bayesianas nos dan ciertas ideas sobre como construirlas. Si consideramos que (regla del producto):

$$P(x_1, x_2, \ldots, x_n) = P(x_n|x_{n-1}, \ldots, x_1)P(x_{n-1}, \ldots, x_1)$$

- Iterando el proceso tenemos que:

$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_n|x_{n-1}, \ldots, x_1)P(x_{n-1}|x_{n-2}, \ldots, x_1) \\ &\quad \cdots P(x_2|x_1)P(x_1) \\ &= \prod_{i=1}^{n} P(x_i|x_{i-1}, \ldots, x_1) \end{aligned}$$

- Esta es la llamada **regla de la cadena**

Construcción de redes bayesianas

- Dadas estas propiedades, podemos afirmar que si $padres(X_i) \subseteq \{X_{i-1}, \ldots, X_1\}$, entonces:

$$P(X_i|X_{i-1}, \ldots, X_1) = P(X_i|padres(X_i))$$

- Esto quiere decir que una red bayesiana es una representación correcta de un dominio sólo si cada nodo es condicionalmente independiente de sus predecesores en orden, dados sus padres
- Para lograr esto, los padres de una variable X_i deben ser aquellos de entre las variables $X_1, \ldots X_{i-1}$ que influyan directamente en X_i

Coste de representación

- Como comentamos, el coste de representar la distribución de probabilidad conjunta de n variables binarias es $O(2^n)$
- La representación de redes bayesianas nos permite una representación mas compacta gracias a la factorización de la distribución conjunta
- Suponiendo que cada nodo de la red tenga como máximo k padres ($k \ll n$), un nodo necesitará 2^k para representar la influencia de sus padres, por lo tanto el espacio necesario será $O(n2^k)$.
- Por ejemplo, con 10 variables y suponiendo 3 padres como máximo tenemos 80 frente a 1024, con 100 variables y suponiendo 5 padres tenemos 3200 frente a aproximadamente 10^{30}

Inferencia en Redes Bayesianas

- El objetivo de la inferencia probabilística será calcular la distribución de probabilidad a posteriori de un conjunto de variables dada la observación de un evento (valores observados para un subconjunto de variables)
- Denotaremos como X la variable sobre la que queremos conocer la distribución
- $\mathbf{E}$ será el conjunto de variables de las que conocemos su valor $E_1, \ldots, E_n$
- $\mathbf{Y}$ será el conjunto de variables que no hemos observado $Y_1, \ldots, Y_n$ (variables ocultas)
- De esta manera $\mathbf{X} = \{X\} \cup \mathbf{E} \cup \mathbf{Y}$ será el conjunto completo de variables
- Nos plantearemos el cálculo de $P(X|\mathbf{e})$

Inferencia Exacta

- **Inferencia por enumeración**: Cualquier probabilidad condicionada se puede calcular como la suma de todos los posibles casos a partir de la distribución de probabilidad conjunta.

$$P(X|\mathbf{e}) = \alpha P(X, \mathbf{e}) = \alpha \sum_{y} P(X, \mathbf{e}, \mathbf{y})$$

- La red bayesiana nos permite factorizar la distribución de probabilidad conjunta y obtener una expresión mas fácil de evaluar
- Usando la red bayesiana ejemplo podemos calcular la probabilidad de ser fumador si se ha tenido un infarto y no se hace deporte

$$P(Fumador|Infarto = si, Deporte = no)$$

Inferencia Exacta: Ejemplo

La distribución de probabilidad conjunta de la red sería:

$$P(D, A, S, F, I) = P(I|S, F)P(F)P(S|D, A)P(D)P(A)$$

Debemos calcular $P(F|I = si, D = no)$, por lo tanto tenemos

$$\begin{aligned} P(F|I = s, D = n) &= \alpha P(F, I = s, D = n) \\ &= \alpha \sum_{A \in \{e, \neg e\}} \sum_{S \in \{a, n\}} P(D = n, A, S, F, I = s) \\ &= \alpha P(D = n)P(F) \sum_{A \in \{e, \neg e\}} P(A) \sum_{S \in \{a, n\}} P(S|D = n, A)P(I = s|S, F) \end{aligned}$$

Inferencia Exacta: Ejemplo

Si enumeramos todas las posibilidades y las sumamos de acuerdo con la distribución de probabilidad conjunta tenemos que:

$$
\begin{aligned}
&P(\mathit{Fumador}|\mathit{Infarto} = \mathit{si}, \mathit{Deporte} = \mathit{no}) \\
&= \alpha\langle 0{,}9 \cdot 0{,}4 \cdot (0{,}4 \cdot (0{,}25 \cdot 0{,}8 + 0{,}75 \cdot 0,6) + 0{,}6 \cdot (0{,}7 \cdot 0{,}8 + 0{,}3 \cdot 0{,}6)) \\
&\qquad 0{,}9 \cdot 0{,}6 \cdot (0{,}4 \cdot (0{,}25 \cdot 0{,}7 + 0,75 \cdot 0{,}3) + 0{,}6 \cdot (0{,}7 \cdot 0{,}7 + 0{,}3 \cdot 0{,}3)\rangle \\
&= \alpha\langle 0{,}253, 0,274\rangle \\
&= \langle 0,48, 0{,}52\rangle
\end{aligned}
$$

Algoritmo de eliminación de variables

- El **algoritmo de eliminación de variables** intenta evitar la repetición de cálculos que realiza la inferencia por enumeración
- El algoritmo utiliza técnicas de programación dinámica de manera que se guardan cálculos intermedios para cada variable para reutilizarlos (*factores*)
- El cálculo de la probabilidad de la pregunta se realiza evaluando la expresión de la distribución de probabilidad conjunta de izquierda a derecha
- Los *factores* correspondientes a cada variable se van acumulando según se necesita
- La ventaja de este algoritmo es que las variables no relevantes desaparecen al ser factores constantes

Algoritmo de eliminación de variables

funcion ELIMINACION-Q(X,**e**,rb) **retorna** distribucion sobre X
  $factores$=[]; $vars$=REVERSE(VARS(rb))
  para cada var **en** $vars$ **hacer**
    $factores$=concatena(factores,CALCULA-FACTOR(var,**e**))
    si var es variable oculta **entonces**
      $factores$=PRODUCTO-Y-SUMA(var,$factores$)
  retorna NORMALIZA(PRODUCTO($factores$))

- CALCULA-FACTOR genera el factor correspondiente a la variable en la función de la distribución de probabilidad conjunta
- PRODUCTO-Y-SUMA multiplica los factores y suma respecto a la variable oculta
- PRODUCTO multiplica un conjunto de factores

Algoritmo de eliminación de variables - Factores

- Un factor corresponde a la probabilidad de un conjunto de variables dadas las variables ocultas
- Se representa por una tabla que para cada combinación de variables ocultas da la probabilidad de las variables del factor

$f_X(Y, Z) =$

Y	Z	
C	C	0.2
C	F	0.4
F	C	0.8
F	F	0.6

- Los factores tienen dos operaciones: suma y producto

Suma de Factores

- La suma se aplica a un factor y sobre una variable oculta del factor. Como restado obtenemos una matriz reducida en la que las filas del mismo valor se han acumulado

$$f_{X\bar{Z}}(Y) = \sum_Z f_X(Y,Z) =$$

Y	
C	0.6
F	1.4

- Es igual que una operación de agregación sobre una columna en bases de datos

Producto de Factores

- El producto de factores permite juntar varios factores entre ellos utilizando las variables ocultas comunes

$$f_{X_1X_2}(Y,W,Z) = f_{X_1}(Y,Z) \times f_{X_2}(Z,W) =$$

Y	Z		Z	W		Y	Z	W	
C	C	0.2	C	C	0.3	C	C	C	$0{,}2 \times 0{,}3$
C	F	0.8	C	F	0.7	C	C	F	$0{,}2 \times 0{,}7$
F	C	0.4	F	C	0.1	C	F	C	$0{,}8 \times 0{,}1$
F	F	0.6	F	F	0.9	C	F	F	$0{,}8 \times 0{,}9$
						F	C	C	$0{,}4 \times 0{,}3$
						F	C	F	$0{,}4 \times 0{,}7$
						F	F	C	$0{,}6 \times 0{,}9$
						F	F	F	$0{,}6 \times 0{,}3$

- Es igual que una operación de join en una base de datos multiplicando los valores de las columnas de datos

Algoritmo de eliminación de variables - ejemplo

Volveremos a calcular $P(Fumador|Infarto = si, Deporte = no)$ a partir de la distribución de probabilidad conjunta:

$$P(D, A, S, F, I) = P(I|S, F)P(F)P(S|D, A)P(D)P(A)$$

Debemos calcular $P(F|I = si, D = no)$, por lo tanto tenemos

$$\begin{aligned} P(F|I = s, D = n) &= \alpha P(I = s, F, D = n) \\ &= \alpha \sum_{A \in \{e, \neg e\}} \sum_{S \in \{a, n\}} P(D = n, A, S, F, I = s) \end{aligned}$$

En esta ocasión no sacamos factores comunes para seguir el algoritmo

$$\alpha P(D = n) \sum_{A \in \{e, \neg e\}} P(A) \sum_{S \in \{a, n\}} P(S|D = n, A)P(F)P(I = s|S, F)$$

Algoritmo de eliminación de variables - ejemplo

El algoritmo empieza calculando el factor para la variable *Infarto* ($P(I = s|S, F)$), esta tiene fijo su valor a `si`, depende de las variables *Presión Sanguinea* y *Fumador*

$f_I(S, F)=$

S	F	
a	s	0.8
a	n	0.7
n	s	0.6
n	n	0.3

La variable fumador ($P(F)$) no depende de ninguna otra variable, al ser la variable que preguntamos el factor incluye todos los valores

$f_F(F)=$

F	
s	0.4
n	0.6

Algoritmo de eliminación de variables - ejemplo

La variable *Presión Sanguinea* ($P(S|D = n, A)$), depende de las variable *Deporte* que tiene fijo su valor a `no` y *Alimentación*. Esta es una variable oculta, por lo que se debe calcular para todos sus valores

$f_S(S, A)=$

S	A	
a	e	0.25
a	¬e	0.7
n	e	0.75
n	¬ e	0.3

Al ser la variable *Presión Sanguinea* una variable oculta debemos acumular todos los factores que hemos calculado

$$f_S(S, A) \times f_F(F) \times f_I(S, F)$$

Algoritmo de eliminación de variables - ejemplo

$f_{FI}(S,F) = f_F(F) \times f_I(S,F) =$

S	F	
a	s	0.8×0.4
a	n	0.7×0.6
n	s	0.6×0.4
n	n	0.3×0.6

$f_{FIS}(S,F,A) = f_{FI}(S,F) \times f_S(S,A) =$

S	F	A	
a	s	e	$0.8 \times 0.4 \times 0.25$
a	s	$\neg$e	$0.8 \times 0.4 \times 0.7$
a	n	e	$0.7 \times 0.6 \times 0.25$
a	n	$\neg$e	$0.7 \times 0.6 \times 0.7$
n	s	e	$0.6 \times 0.4 \times 0.75$
n	s	$\neg$e	$0.6 \times 0.4 \times 0.3$
n	n	e	$0.3 \times 0.6 \times 0.75$
n	n	$\neg$e	$0.3 \times 0.6 \times 0.3$

Algoritmo de eliminación de variables - ejemplo

Y ahora sumamos sobre todos los valores de la variable S para obtener el factor correspondiente a la variable *Presión Sanguinea*

$$f_{FI\overline{S}}(F, A) = \sum_{S \in \{a,n\}} f_{FIS}(S, F, A) =$$

F	A	
s	e	$0.8 \times 0.4 \times 0.25 + 0.6 \times 0.4 \times 0.75 = 0.26$
s	¬e	$0.8 \times 0.4 \times 0.7 + 0.6 \times 0.4 \times 0.3 = 0.296$
n	e	$0.7 \times 0.6 \times 0.25 + 0.3 \times 0.6 \times 0.75 = 0.24$
n	¬e	$0.7 \times 0.6 \times 0.7 + 0.3 \times 0.6 \times 0.3 = 0.348$

Algoritmo de eliminación de variables - ejemplo

El factor de la variable *Alimentación* ($P(A)$) no depende de ninguna variable, al ser una variable oculta generamos todas las posibilidades

$f_A(A)=$

F	
e	0.4
¬e	0.6

Ahora debemos acumular todos los factores calculados

$f_{AFI\overline{S}}(A) = f_A(A) \times f_{FI\overline{S}}(F,A)=$

F	A	
s	e	$0.26 \times 0.4 = 0.104$
s	¬e	$0.296 \times 0.6 = 0.177$
n	e	$0.24 \times 0.4 = 0.096$
n	¬e	$0.348 \times 0.6 = 0.208$

Algoritmo de eliminación de variables - ejemplo

Y ahora sumamos sobre todos los valores de la variable A para obtener el factor correspondiente a la variable *Alimentación*

$$f_{\overline{AFI}\overline{S}}(F) = \sum_{A \in \{e, \neg e\}} f_{AFI\overline{S}}(A) =$$

F	
S	$0.104 + 0.177 = 0.281$
n	$0.096 + 0.208 = 0.304$

Y por último la variable *Deporte* ($P(D = n)$) tiene el valor fijado a *no* y dado que no depende de la variable *fumador* se puede obviar, ya que es un factor constante.
Ahora, si normalizamos a 1

$$P(F|I = s, D = n) =$$

F	
S	0.48
n	0.52

Complejidad de la inferencia exacta

- La complejidad del algoritmo de eliminación de variables depende del tamaño del mayor factor, que depende del orden en el que se evalúan las variables y la topología de la red
- El orden de evaluación que escogeremos será el topológico según el grafo
- La complejidad de la inferencia exacta es NP-hard en el caso general
- Si la red bayesiana cumple que para cada par de nodos hay un único camino no dirigido (**poliárbol**) entonces se puede calcular en tiempo lineal
- Para obtener resultados en el caso general se recurre a algoritmos aproximados basados en técnicas de muestreo