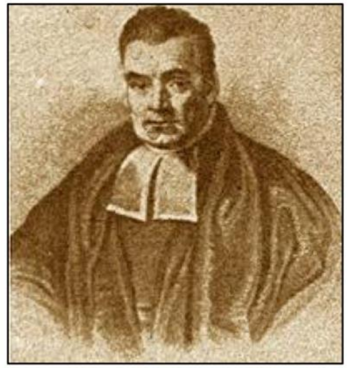
Redes Bayesianas

Introducción

Regla de Bayes

- **Propuesta en 1763 por el Reverendo T. Bayes**
  - $P(A|B) = P(B|A)\,P(A) \;/\; P(B)$
  - Es una consecuencia de la regla del producto:
    - $P(A|B)P(B) = P(A,B) = P(B|A)P(A)$



Thomas Bayes

- **De forma intuitiva:**
  - La probabilidad de una hipótesis A dada una evidencia B: P(A|B) es proporcional a probabilidad de la hipótesis P(A) multiplicada por el grado en que la hipótesis predice los datos P(B|A)
- **Aplicabilidad**
  - En muchos problemas dado un conjunto de datos (evidencia) B tenemos que seleccionar la hipótesis A más probable mediante P(A|B)

Introducción

Forma general del Teorema de Bayes

- **Forma general de la Regla de Bayes**
  - Si se tiene un conjunto de proposiciones $\{A_1, A_2, \ldots, A_m\}$ completas y mutuamente excluyente se tiene:

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{P(B|A_1)\,P(A_1) + \ldots\, P(B|A_n)\,P(A_m)}$$

O lo que es lo mismo, si tiene una variable aleatoria A con valores $a_1, a_2, \ldots, a_m$

$$P(a_i|B) = \frac{P(B|a_i)\,P(a_i)}{P(B|a_1)\,P(a_1) + \ldots\, P(B|a_n)\,P(a_m)}$$

Introducción

Ejemplo

- **Intentemos resolver un caso real con probabilidades:**
  - Se pretende determinar si un alumno conoce un concepto en base a la resolución de un ejercicio.
    - En este caso:
      - Hipótesis (SC): Sabe_Concepto (variable aleatoria con dos valores verdadero y falso)
      - Evidencia (RE): Resuelve_Ejercicio (variable aleatoria con dos valores positivo y negativo)
    - Aplicando la Regla de Bayes:

      $P(sc|re)= P(re|sc) \quad P(sc) \quad /(P(re|sc) \quad P(sc) \quad + P(re| \neg sc) \quad P(\neg sc))=0.95$

      $P(\neg sc |re)=0.05$
    - Al elegir la hipótesis más probable debemos concluir que si resuelve el ejercicio sabe el concepto

Introducción

Ejemplo

- **Continuamos con el ejemplo:**
  - ¿Y si hay varios ejercicios $E_1,\ldots, E_m$?
    - Supondremos que cada ejercicio RE_1, $RE_2,\ldots$, RE_m es una variable aleatoria que indica si se resuelve con dos valores: verdadero y falso.
  - Entonces si queremos calcular la probabilidad de que el alumno sepa el concepto necesitamos calcular:

    $P(SC|\ E_1,RE_2,\ldots,RE_m)=P(RE_1,,\ldots,RE_m|SC)P(SC)/P(RE_1,RE_2,\ldots,RE_m)$

  - Si al alumno se le hace un conjunto de 7 ejercicios:
    - Entonces para almacenar la tabla de probabilidad conjunta $P(RE_1,RE_2,\ldots,RE_m|\ SC)$ se necesitan guardar unos 2^7 números reales
    - ¿De donde sacamos los números ?
    - ¿Cómo hacemos los cálculos computacionalmente eficientes?

Introducción

- **Independencia**
  - Decimos que dos proposiciones A_1 y A_2 son independientes si el conocimiento de una no cambia la probabilidad de la otra
    - Por ejemplo si
      - A_1="Es rubio" , A_2="Tiene la piel clara" ,A_3="Lloverá mañana"
      - A_1 y A_3 son independientes A_1 y A_2 no.
  - Formalmente A_1,A_2 son independientes si $P(A_1|A_2)=P(A_1)$
    o de forma equivalente: $P(A_2|A_1)=P(A_2)$
    o utilizando la regla del producto $P(A_1 \wedge A_2)= P(A_1)\ P(A_2)$
  - Entonces $P(A_1 \wedge A_2 \wedge \ldots \wedge A_n)= P(A_1)\ P(A_2) \ldots P(A_n)$
    Para especificar la distribución conjunta de n variables se necesitan o(n) números en lugar de $o(2^n)$
  - Dos variables aleatorias son independientes si el conocimento del valor que toma una no cambia la probabilidad de los valores de la otra: $P(A_1=c|A_2=d) = P(A_1=c)$

Introducción

- **Pero...**
  - La condición de independencia es muy restrictiva.
  - Por ejemplo, los resultados de los ejercicios en la enseñanza no suelen ser independientes.
- **Independencia condicional**
  - Se dice que dos proposiciones A_1,A_2 son independientes dada una tercera B si cuando B está presente el conocimiento de una no influye en la probabilidad de la otra: $P(A_1|A_2,B)=P(A_1|B)$

    o de forma equivalente: $P(A_2|A_1,B)=P(A_2|B)$

    o de forma equivalente: $P(A_1 \wedge A_2 |B)= P(A_1|B)\ P(A_2|B)$
    - Ejemplo:
      - A_1="Tengo congestión nasal" A_2="Tengo fiebre" A_3="Tengo gripe"
      - A_1 y A_2 son dependientes pero son independientes si se conoce A_3.
  - Ahora se tiene: $P(A_1 \wedge A_2 \wedge \ldots \wedge A_n |B)=P(A_1|B)\ P(A_2|B)\ \ldots\ P(A_n|B)$
    - Tenemos o(n) números en lugar de o(2^n)

Introducción

Independencia condicional

- **Finalizamos el ejemplo:**
  - ¿Y si hay varios ejercicios $E_1, E_2, \ldots, E_m$?
  - Como vimos, para calcular la probabilidad de que el alumno sepa el concepto necesitamos calcular:

    $$P(SC| E_1, RE_2, \ldots, RE_m) = P(RE_1,, \ldots, RE_m | SC) P(SC) / P(RE_1, RE_2, \ldots, RE_m)$$

    Si los resultado de los ejercicios $E_1, E_2, \ldots, E_m$ son independientes dado el concepto (aproximación que suele dar buenos resultados):

    $$P(RE_1,, \ldots, RE_m | SC) = P(RE_1 | SC)\ P(RE_2 | SC)\ \ldots\ P(RE_m | SC)$$
  - El problema a resolver ya es abordable:

Introducción

La clave para hacer factible el cálculo de probabilidades es la introducción explícita de la independencia entre variables

El modelo más extendido de representación de independencias lo constituyen las Redes Bayesianas, también llamadas Redes de creencia, Redes probabilísticas, Redes causales o Mapas de conocimiento

Introducción

Redes Bayesianas

1) Es una **estructura de datos para representación de conocimiento incierto**

2) Representa la **dependencia entre variables**, y especifica en forma concisa la distribución de probabilidad conjunta

3) Es una **representación gráfica**

Generalmente suele ser fácil especificar qué relaciones de dependencia condicional se dan, esto implica

- Determinar la **topología de la red**
- Especificar las **probabilidades condicionales** de los nodos con dependencia directas
- Calcular cualquier otro valor de probabilidad

Introducción

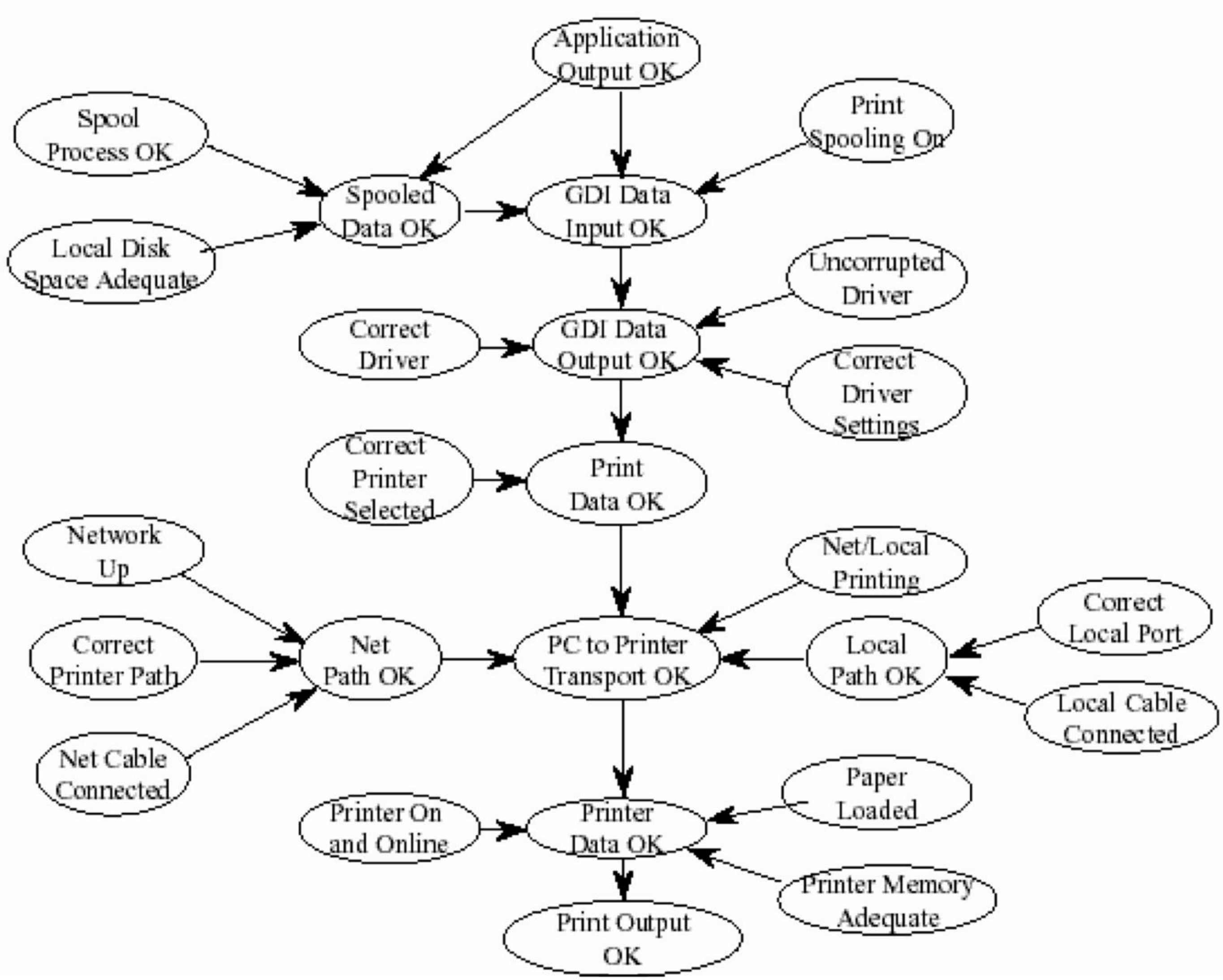
Una **red bayesiana** es:

Un conjunto de **nodos** que representan variables o entidades del mundo real

Un conjunto de **enlaces** que representan relaciones de influencia causal entre los nodos

Una serie de **parámetros** (probabilidades condicionadas de cada nodo dados sus padres) que cuantifican la relación entre los nodos.

Introducción



Diagnóstico de Problemas de Impresión (Heckerman)

Introducción

Una conexión tipo A → B **indica dependencia** o relevancia directa entre las variables.

En este caso se está representando que B depende de A o que A es la causa de B y B el efecto de A. También se dice que A es padre (o madre) de B y que B es el hijo (o hija) de A.

Aunque la presencia de arcos entre nodos codifica información esencial sobre el modelo representado en la red, la **ausencia de arcos** entre nodos aporta una valiosa información ya que el grafo codifica **independencia condicional**.

Introducción

En una RB, la información proporcionada por una o más variables que se observan (**evidencia**) se propaga por la red y actualiza nuestra creencia acerca de las variables no observadas. A este proceso se le llama **inferencia**.

Es posible aprender las probabilidades condicionales que describen las relaciones entre las variables a partir de los datos. Incluso es posible aprender la estructura completa de la red a partir de datos completos o con algunos de sus valores desconocidos.

Las RB pueden utilizarse para tomar decisiones óptimas introduciendo posibles acciones y la utilidad de sus resultados

Introducción

Resumiendo, en una RB

- Cada nodo de la red representa una variable aleatoria
- Un arco del nodo X al nodo Y, significa que la variable X tiene una influencia directa sobre Y. Decimos que X es el padre (madre) de Y.
- Cada nodo X tiene una tabla de probabilidad condicional que cuantifica el efecto que los padres de X tienen sobre X
- Es un <u>grafo dirigido acíclico</u> (GDA)

Introducción

Algunas aplicaciones de las RB en empresas

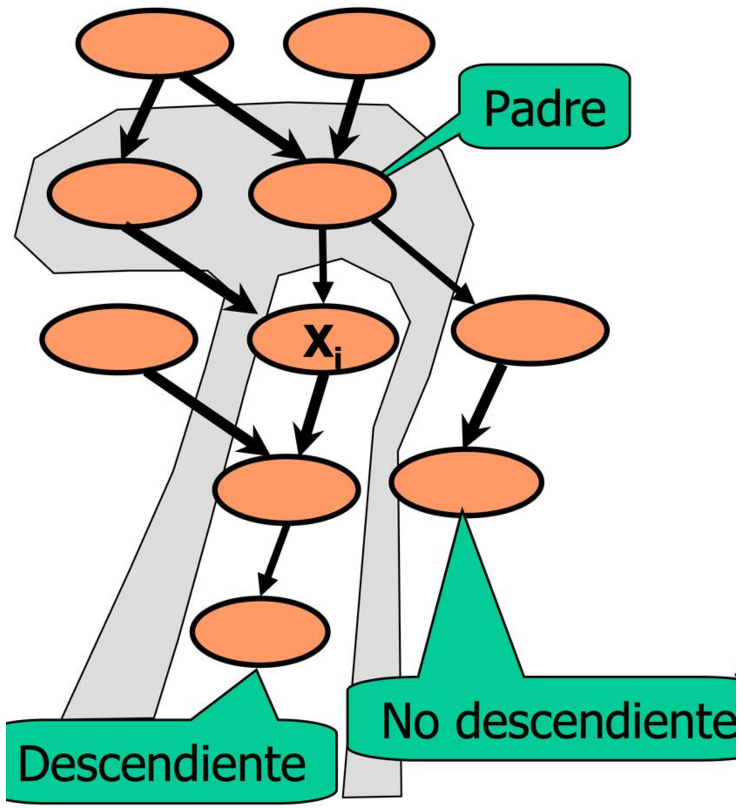
– Microsoft: AnswerWizard (Office); Diagnóstico de problemas de usuario (Aladdin); HomeHealth en la red de Microsoft (MSN)
– Intel: Diagnóstico de fallos de procesadores
– HP: Diagnóstico de problemas de impresora
– Nokia: Diagnóstico de redes celulares
– Nasa: Sistema de ayuda a la decisión en misiones espaciales

Introducción

Dada una RB con nodos X1, X2,...,Xn. Si Padres(Xi) son los padres de Xi y NoDescendientes(Xi) son los nodos que no son descendientes de Xi.

Entonces para cada variable Xi se tiene que esta es independiente de sus NoDescendientes, dados sus Padres. Es decir,

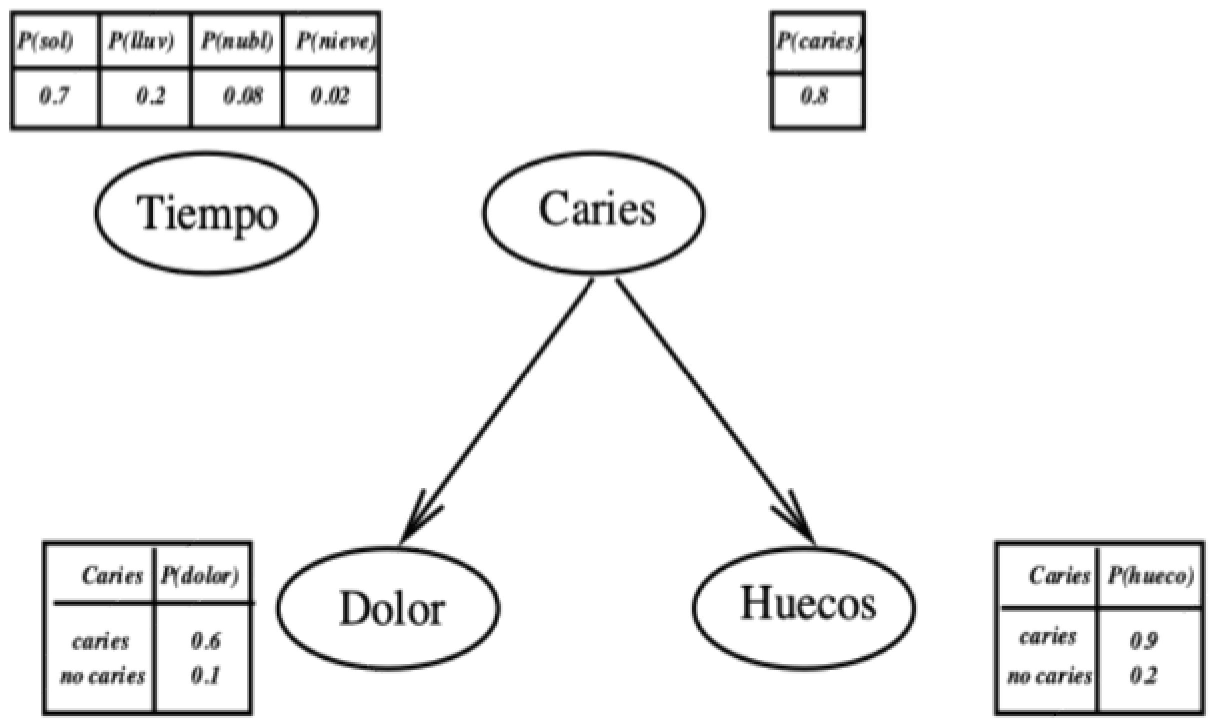
Indep(Xi; NoDescendientes(Xi) | Pa(Xi))



Introducción

Ejemplo

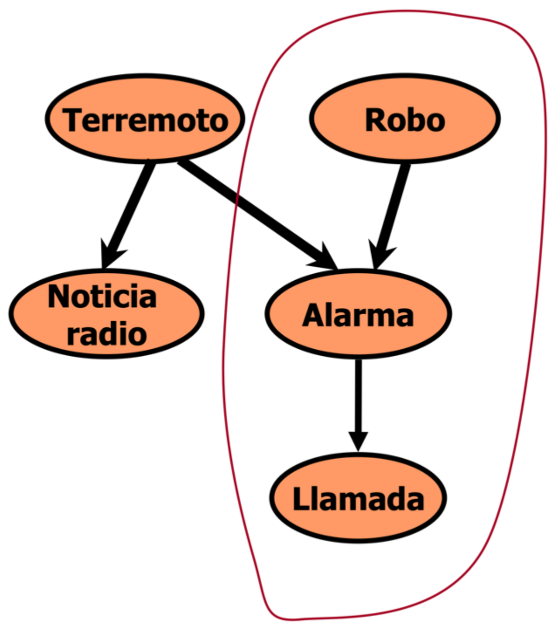
- Caries es una causa directa de Dolor y Huecos
- Dolor y Huecos son condicionalmente independientes dada Caries
- Tiempo es independiente de las restantes variables



Introducción

R y L son dependientes:
- Si hay un robo es más probable suene la alarma, lo que hace más probable que reciba una llamada.
- Si recibo una llamada se incrementa la probabilidad de que haya sonado la alarma y por tanto de que me hayan robado.

R y L son independientes si se conoce A:
- Si hay un robo ya no es más probable que suene la alarma (ya se sabe si suena o no)
- Si recibo una llamada ya no se incrementa la probabilidad de que suene la alarma (ya se sabe si suena o no).



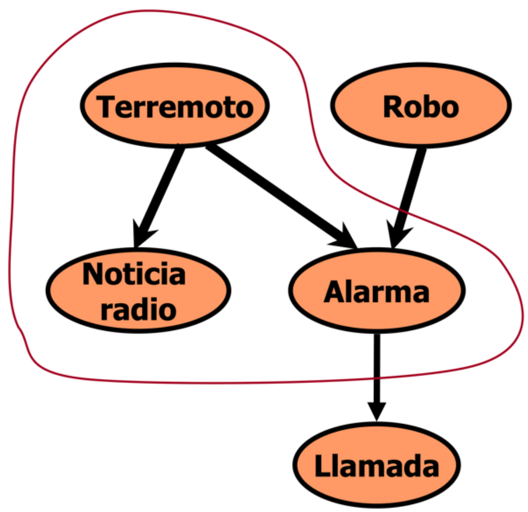
L es independiente de sus no-descendientes

T,R,N, dados sus padres A

Introducción

N y A son dependientes:
- Si oigo en la radio que ha habido un terremoto es más probable que éste haya ocurrido, lo que hace más probable que suene la alarma.
- Si suena la alarma se incrementa la probabilidad de que haya ocurrido un terremoto y por tanto de que oiga la noticia en la radio.

N y A son independientes si se conoce T:
- Si oigo en la radio que ha habido un terremoto ya no es más probable que éste haya ocurrido (ya se sabe si ha ocurrido o no).
- Si suena la alarma ya no se incrementa la probabilidad de que haya ocurrido un terremoto (ya se sabe si ocurrió)



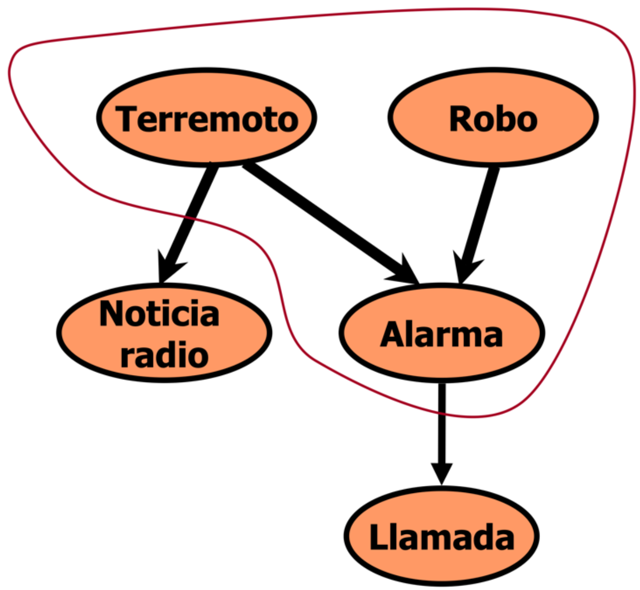
N es independiente de sus no-descendientes

R,A,L dados sus padres T

Introducción

T y R son dependientes si se conoce A:
- Si suena la alarma y ocurre una de las causas (terremoto) me creo menos la otra (robo)
- Si suena la alarma y ocurre una de las causas (robo) me creo menos la otra (terremoto)
- A este efecto se le llama “eliminación de explicaciones”

T y R son independientes:
- Si desconozco si suena la alarma y ocurre una de las causas (terremoto) no hay razón para creer menos la otra (robo)
- Si desconozco si suena la alarma y ocurre una de las causas (robo) no hay razón para creer menos la otra (terremoto)

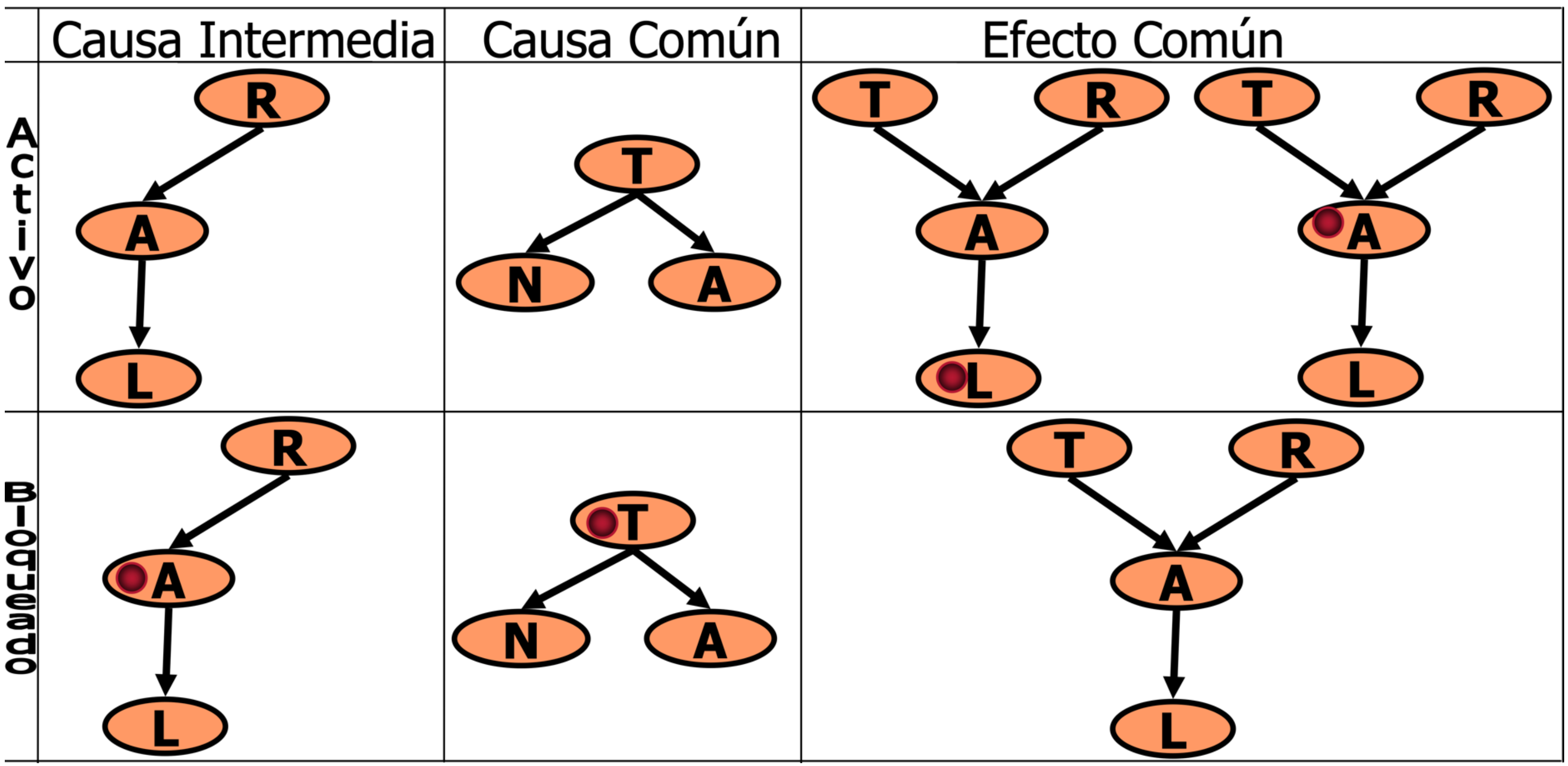


T es independiente de sus no-descendientes R dados sus padres (ninguno).

Transmisión de Información en la red

Un camino del grafo puede estar:

- Activo: si pasa información por el
- Bloqueado: si no pasa



Transmisión de Información en la red

Las redes bayesianas son una **representación gráfica del principio de independencia condicional** en términos probabilísticos.

Este principio dice: sean tres conjuntos X, Y y Z de variables; diríamos que los conjuntos X e Y son (condicionalmente) independientes dado el conjunto Z si y solamente si

$$P(x \mid z) = P(x \mid y, z)$$

Dicho de otro modo, dos variables X e Y, son independientes en términos probabilísticos de una tercera Z si y solamente si

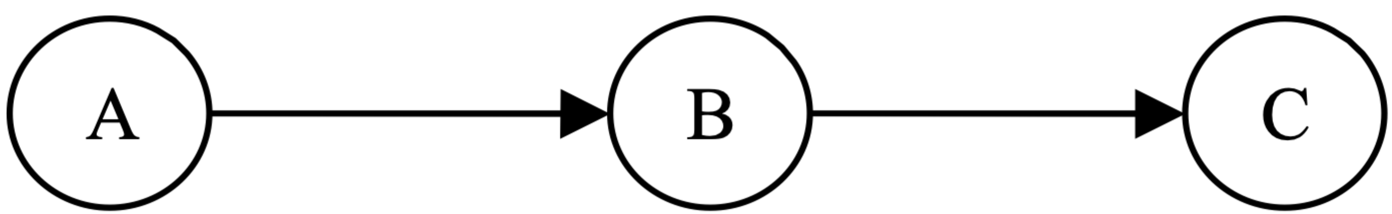
$$P(xy \mid z) = P(x \mid z) \times P(y \mid z)$$

La consecuencia fundamental de este principio es que la probabilidad de X es la misma condicionándola a Z o condicionándola a Z e Y.

Transmisión de Información en la red

Cualquier red bayesiana puede descomponerse en **tres tipos de conexiones básicas**, cada una con propiedades diferentes en la propagación de probabilidades.

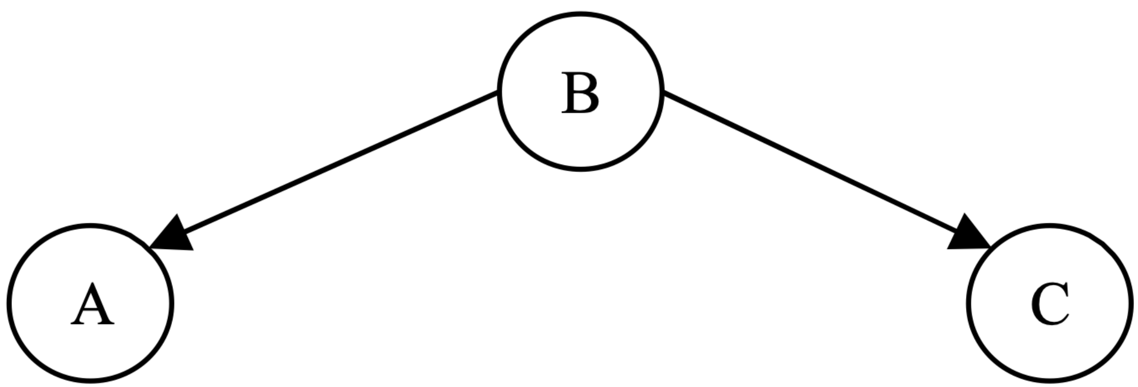
1) En primer lugar, las conexiones seriales o cadenas causales representan un conjunto de variables asociadas linealmente que denota dependencia entre las variables. La variable B depende de A y la variable C depende del valor de B. Cuando sabemos algo sobre A podemos modificar nuestra creencia sobre el estado de B y esta información se propagará hasta C. Sin embargo, si encontramos una evidencia sobre B, añadir evidencias sobre C no alterará nuestro conocimiento sobre A y viceversa. En este caso decimos que A y C son condicionalmente independientes dado B.



Transmisión de Información en la red

Cualquier red bayesiana podría descomponerse en **tres tipos de conexiones básicas**, cada una con propiedades diferentes en el proceso de propagación de probabilidades.

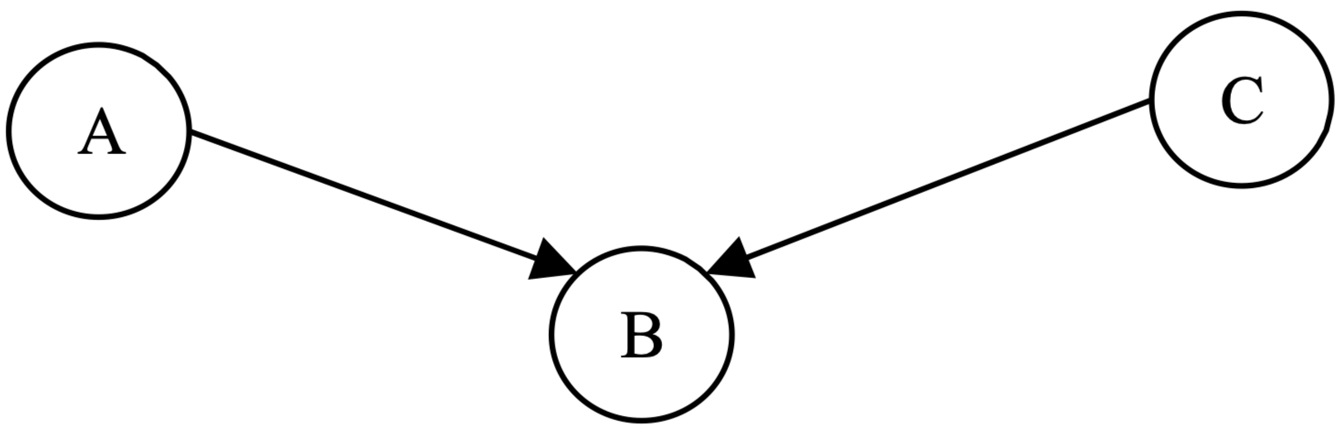
2) En segundo lugar, las conexiones divergentes, también conocidas como clasificadores ingenuos de Bayes. Tenemos un nodo padre (o clase) que proyecta sus arcos sobre varios hijos. Cuando no conocemos el estado de la variable madre existe dependencia entre las variables hijas. Sin embargo, cuando el estado de esta variable se conoce, las evidencias sobre las variables hijas no se propagarán entre ellas. En nuestro ejemplo diremos que A y C son independientes dado B.



Transmisión de Información en la red

Cualquier red bayesiana podría descomponerse en **tres tipos de conexiones básicas**, cada una con propiedades diferentes en el proceso de propagación de probabilidades.

3) En tercer lugar, las conexiones convergentes. En ellas, varias variables apuntan con sus arcos hacia una variable de convergencia. En este tipo de conexiones las variables madre son independientes entre si. Sin embargo, tenemos una evidencia sobre la variable hija, las variables madre se tornarán dependientes. En nuestro ejemplo diríamos que A y C son condicionalmente dependientes dado B.



Resumen de las propiedades de independencia condicional

Independencia a priori de los nodos que no tienen ningún antepasado común

Independencia condicional de los nodos hermanos con respecto a su padre

Independencia condicional entre un nodo y los antepasados de sus padres

Definición formal de Red Bayesiana

Una **red bayesiana** es:

- Un conjunto **exhaustivo y excluyente** de *variables* proposicionales, *V*
- Un conjunto *E* de ***relaciones binarias*** definidas sobre las variables de *V*
- Una ***distribución de probabilidad conjunta*** *P* definida sobre las variables de *V*,

tales que

(*V*, *E*) es un grafo ***acíclico*, *conexo* y *dirigido*** *G*.

(*G*, *P*) cumple las ***hipótesis de independencia condicional***

$\forall X \in V$ *y* $\forall \boldsymbol{Y} \subseteq V - \{X \cup de(X)\} \quad P(X/pa(X), Y) = P(X/pa(X)),$

con *de* = descencientes y *pa* = padres

cada variable es condicionalmente independiente de sus anteriores, dados sus padres en la red

Dimensión Cuantitativa

Existen **tres elementos esenciales que caracterizan la dimensión cuantitativa de una red bayesiana**

1) El concepto de probabilidad como un grado de creencia subjetiva relativa a la ocurrencia de un evento.

2) Un conjunto de funciones de probabilidad condicionada que definen a cada variable en el modelo

3) El teorema de Bayes como herramienta básica para actualizar probabilidades con base en experiencia.

Dimensión Cuantitativa

Tenemos, como mínimo, cuatro formas de entender la probabilidad: desde un punto de vista clásico, desde una perspectiva empírica, axiomáticamente y la concepción bayesiana o subjetiva.

De una forma u otra, **la probabilidad es una manera de cuantificar la incertidumbre asociada a la ocurrencia de eventos** y las redes bayesianas se basan en una idea subjetiva de la probabilidad, siendo el teorema de Bayes el motor de actualización de probabilidades.

Dimensión Cuantitativa

En una red bayesiana cada variable es entendida como una función de probabilidad condicionada a los valores que toman las variables de las que depende

Así pues, **para cada variable**, tenemos que especificar las distribuciones condicionales de Xi dados sus padres Padres(Xi). Si entendemos que esta densidad es P(Xi | Padres(Xi)), entonces la probabilidad global conjunta se deriva de

$$P(x_1,..,x_n) = \prod_{i=1..n} P(x_i \mid \mathrm{Padres}(x_i))$$

Teorema de Factorización

Una RB proporciona una descripción completa del dominio, de forma que:

cualquier elemento de la Distribución de Probabilidad Conjunta se puede calcular a partir de la red

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

$$P(x_1,\ldots,x_n) = P(x_n|x_{n-1},\ldots,x_1)P(x_{n-1}|x_{n-2},\ldots,x_1)\ldots P(x_2|x_1)P(x_1)$$

$$P(x_1,..,x_n) = \prod_{i=1..n} P(x_i \mid x_{i-1},..,x_1)$$

$$\mathrm{P}(\mathrm{x}_1,..,\mathrm{x}_\mathrm{n}) = \prod_{\mathrm{i}=1..\mathrm{n}} \mathrm{P}(\mathrm{x}_\mathrm{i} \mid \mathrm{Padres}(\mathrm{x}_\mathrm{i}))$$

si $\mathrm{Padres}(X_i) \subseteq \{ x_{i-1},\ldots,x_1\}$

Factorización: consecuencias

- Se puede describir P utilizando probabilidades condicionales “locales”

- Si G es un grafo disperso, es decir el número de padres de cada variable está acotado: $|Pa(Xi)| \leq k$ con k un número “pequeño” se obtiene:
  - Representación compacta: El número de parámetros para describir la función de distribución conjunta es lineal en el número n de variables aleatorias o(n). Nótese que el número de parámetros requerido en general es de orden o(2n)
  - Representación modular: Añadir una nueva variable no obliga a actualizar todos los parámetros de la representación

Factorización: consecuencias

Una RB es más compacta que la distribución de probabilidad conjunta correspondiente permitiendo manejar muchas evidencias sin el problema del crecimiento exponencial

Si las variables de una RB reciben influencias directas de un promedio de *k* variables, y hay un total de N variables booleanas, entonces la RB queda especificada por $N.2^k$ probabilidades frente a 2^N números de una distribución de probabilidad completa

Por ejemplo para N=30 y k=5, esto supone 960 números frente a 2^{30}

Factorización

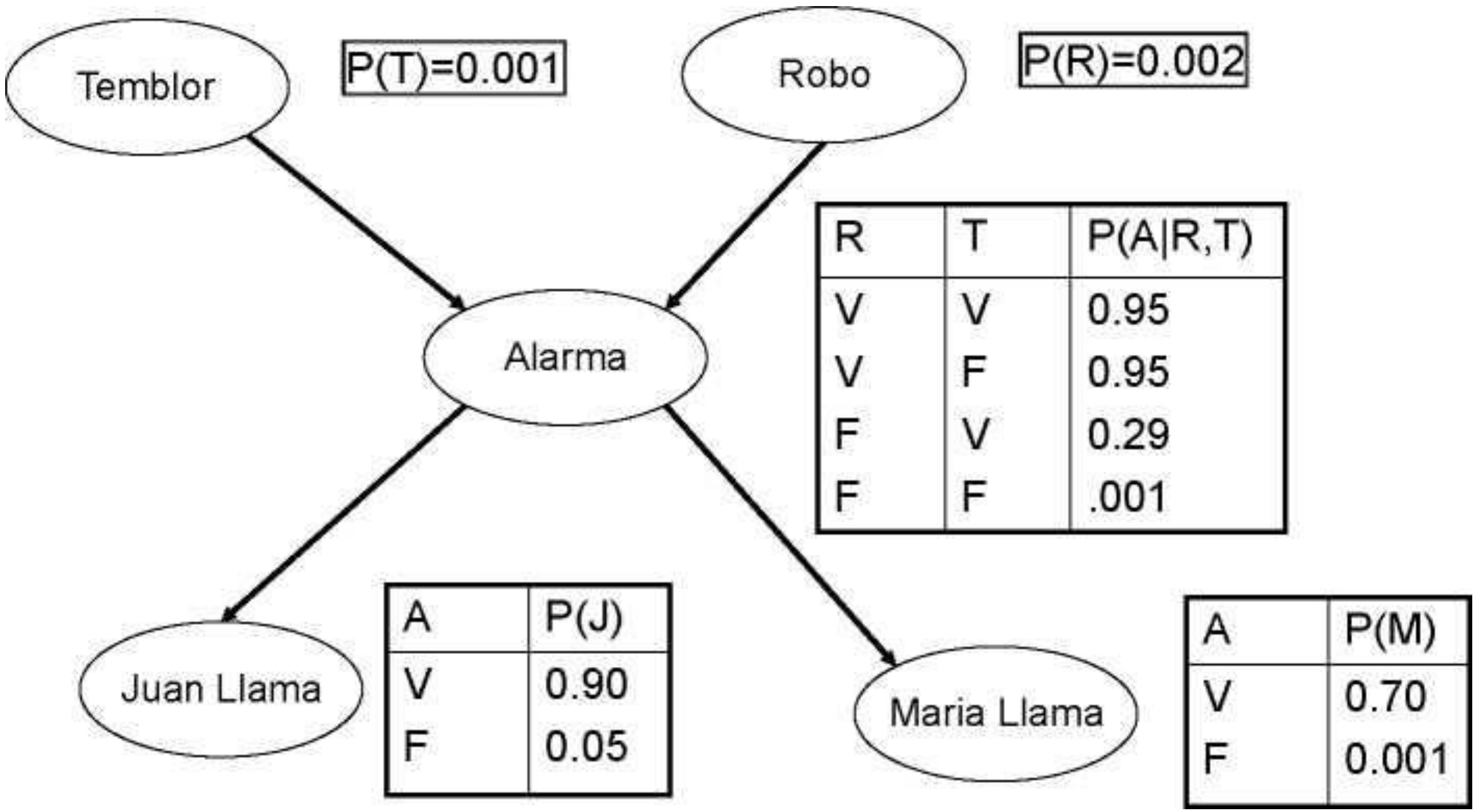
Ejemplo

Tenemos una alarma antirrobo instalada en una casa
- La alarma salta normalmente con la presencia de ladrones
- Pero también cuando ocurren pequeños temblores de tierra

Tenemos dos vecinos en la casa, Juan y María, que han prometido llamar a la policía si oyen la alarma
- Juan y María podrían no llamar aunque la alarma sonara
- Incluso podrían llamar aunque no hubiera sonado

Factorización

Ejemplo



R	T	P(A|R,T)
V	V	0.95
V	F	0.95
F	V	0.29
F	F	.001

A	P(J)
V	0.90
F	0.05

A	P(M)
V	0.70
F	0.001

Factorización

Ejemplo

La topología de la red expresa que

- Robo y Terremoto son causas directas para Alarma
- También, Robo y Temblor son causas para Juan llama y para María llama, pero esa influencia sólo se produce a través de Alarma. Es decir, ni Juan ni María detectan directamente el robo ni los pequeños temblores de tierra
- En la red no se hace referencia directa, por ejemplo, a las causas por las cuales María podría no oír la alarma. Están implícitas en la tabla de probabilidades $P(M \mid A)$

La red expresa que $P(M \mid J, A, T, R) = P(M \mid A)$

Factorización

Ejemplo

Calcula la probabilidad de que la Alarma suene, no haya robo ni Terremoto y que Juan y María llamen.

$P(A, \neg R, \neg T, J, M)$?

$P(A | \neg R, \neg T) P(\neg R) P(\neg T) P(J|A) P(M|A) =$

$$= 0{,}001 \times 0{,}998 \times 0{,}999 \times 0{,}9 \times 0{,}7$$

$$= 0{,}00062$$

$\boxed{P(R | J, \neg M, T) \text{ ?}}$

$$P(x_1, .., x_n) = \prod_{i=1..n} P(x_i \mid Padres(x_i))$$

Factorización

Ejemplo

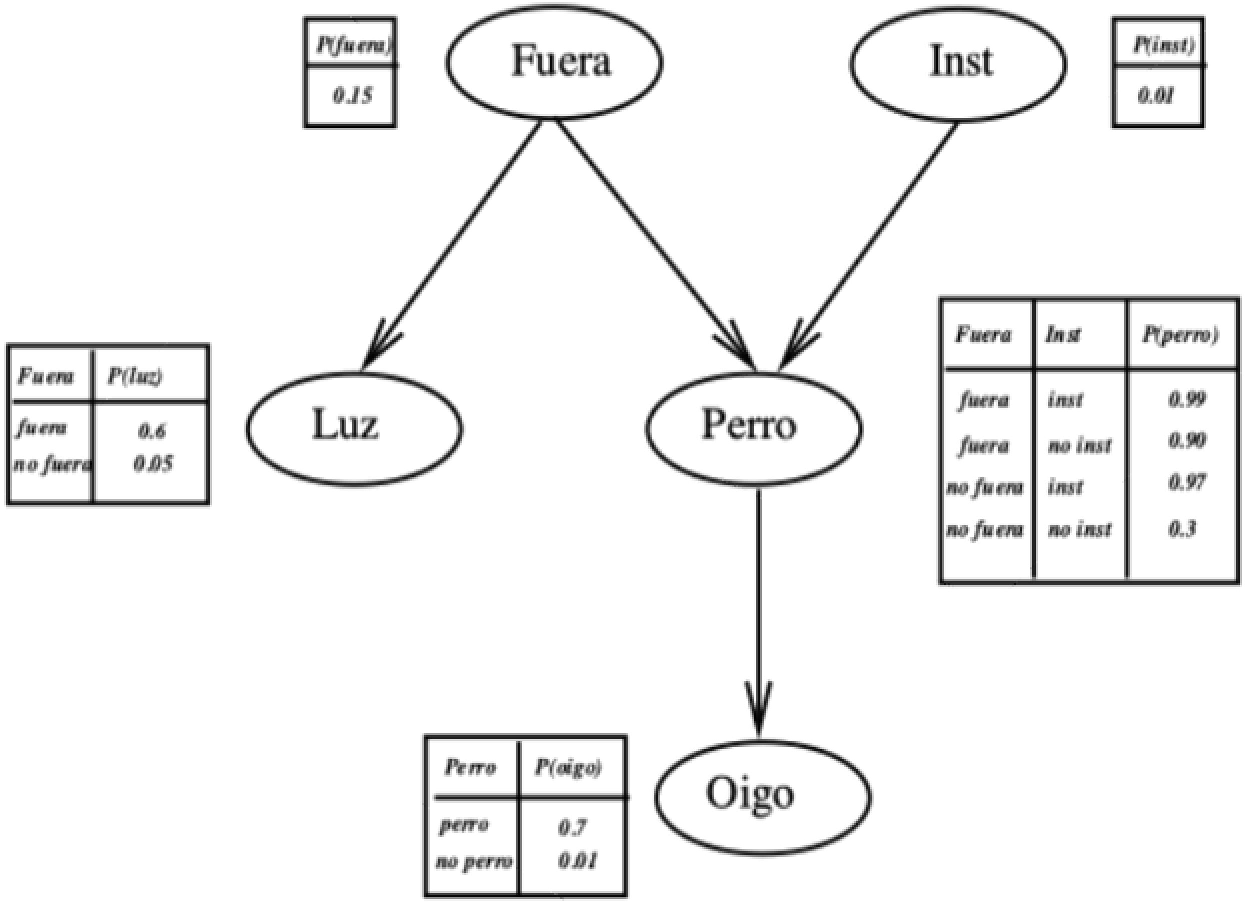
Supongamos que quiero saber si alguien de mi familia está en casa, basándome en la siguiente información:

- Si mi esposa sale de casa, usualmente (pero no siempre) enciende la luz de la entrada.
- Hay otras ocasiones en las que también enciende la luz de la entrada.
- Si no hay nadie en casa, el perro está fuera.
- Si el perro tiene problemas intestinales, también se deja fuera.
- Si el perro está fuera, oigo sus ladridos.
- Podría oír ladridos y pensar que son de mi perro, aunque no fuera así.

Variables aleatorias (booleanas) en este problema:
1) Fuera (nadie en casa), 2) Luz (luz en la entrada), 3) Perro (perro fuera), 4) Inst (problemas intestinales en el perro) y 5) Oigo (oigo al perro ladrar)

Factorización

Ejemplo



Calcular la probabilidad de Oigo, no Inst, no Fuera, Luz y no Perro

Propiedades de independencia condicional

Una RB **es representación correcta** del dominio si cada nodo es independiente respecto de antepasados de sus padres

Conviene escoger a los padres de manera que se satisfaga la condición anterior

En el ejemplo, no hay relación directa entre el hecho de que Maria o Juan llamen y el que se produzca un terremoto o un robo. La relación está mediada por el hecho de que suene la alarma

- $P(M \mid J,A,T,R) = P(M \mid A)$
- $P(J \mid J,A,T,R) = P(J \mid A)$

Metodología de Construcción de una RB

Supongamos dado un conjunto de variables aleatorias que representan un dominio de conocimiento (con incertidumbre)

1. Escoger conjunto de variables X1,…,Xn que describen un dominio.
2. Definir un orden parcial para el conjunto de variables; primero los nodos causales y luego los nodos efecto.
3. Mientras queden variables
   a) Escoger siguiente variable Xi y añadir nodo para ella.
   b) Asignar Padres(Xi) como el conjunto mínimo de nodos X1,…,Xi-1 presente en la red, de manera que sea satisfecha la propiedad de independencia: Ind(Xi ; {X1,…,Xi-1 } - Padres(Xi) | Padres(Xi))
   c) Elaborar la tabla de probabilidad condicional de Xi

Metodología de Construcción de una RB

```
FUNCION CONSTRUYE_RED(VARIABLES)

1. Sea (X_1,...X_n) una ordenación de las variables de VARIABLES
2. Sea RED una red bayesiana ``vacía''
3. PARA i=1,...,n HACER
   3.1 Añadir un nodo etiquetado con X_i a RED
   3.2 Sea padres(X_i) un subconjunto minimal de {X_{i-1},...,X1}
       tal que existe una independencia condicional entre X_i y cada
       elemento de {X_{i-1},...,X1} dado padres(X_i)
   3.3 Añadir en RED un arco dirigido entre cada elemento de padres(X_i)
       y X_i
   3.4 Asignar al nodo X_i la tabla de probabilidad P(X_i|padres(X_i))
4. Devolver RED
```

Metodología de Construcción de una RB

Este método **garantiza la obtención de redes acíclicas**

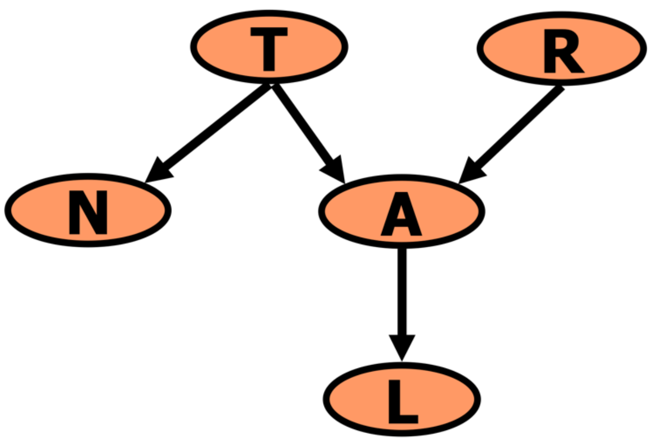
Evita la redundancia en la definición de probabilidades

Evita que se violen los axiomas de probabilidad

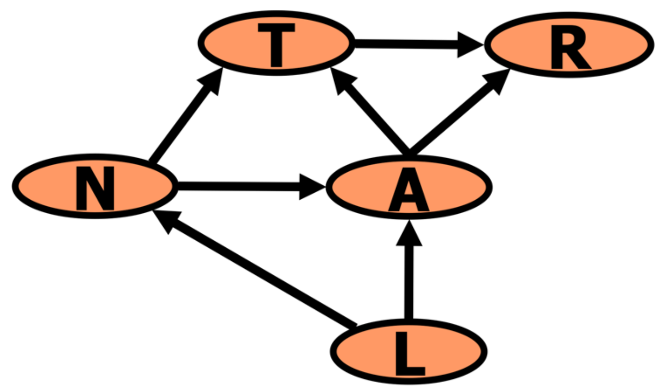
Metodología de Construcción de una RB

Problema: La red resultante depende de la elección del orden entre variables

Orden:T, R, A, L, N



Orden: L, N, A, T, R



Un orden malo puede llevar a representaciones poco eficientes

Metodología de Construcción de una RB

Solución:

- En general, deberíamos empezar por las “causas originales”, siguiendo con aquellas a las que influencian directamente, etc..., hasta llegar a las que no influyen directamente sobre ninguna (modelo causal). Generalmente se puede asumir que los grafos generados a partir de relaciones causales cumplen las condiciones de independencia.

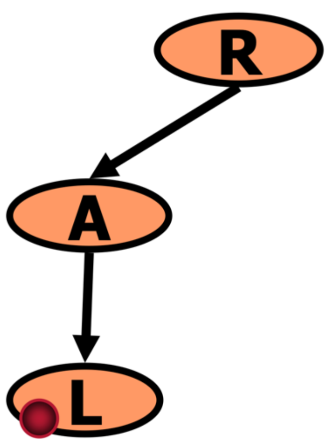
- Esto hará que las tablas reflejen probabilidades “causales” más que “diagnósticos”, lo cual suele ser preferible por los expertos

Inferencia en una RB

Se pretende **hallar la distribución de probabilidad de determinadas variables de interés dados los valores de otras variables que se observan (evidencias)**

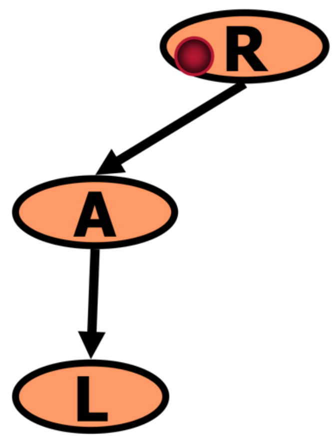
Principales tipos de Inferencias

- Diagnóstico



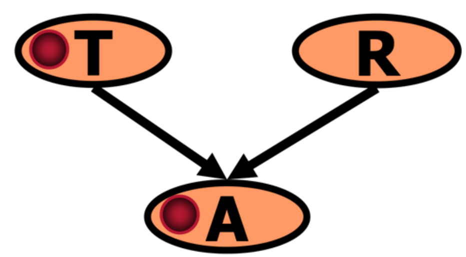
P(robo | llama)

Predicción



P(llama | robo)

Intercausal



P(robo | alarma,¬terremoto)

Tipos de Inferencia

Utilizando el ejemplo de la alarma

Modelo diagnóstico: efectos (síntomas) causas (diagnóstico)

- P(R|J), P(R|J,M)

Modelo predicción: Causas efectos - P(J|R), P(M,R)

Inferencias intercausales: entre las causas de un efecto común

- P(R|A,T)

Inferencias mixtas: combinación de las anteriores

- P(A|J,-T), P(R|J,-T)

Inferencia en una RB

La Inferencia es el efecto de la evidencia propagada por la red para conocer probabilidades a posteriori

La Propagación se obtiene dando valores a ciertas variables (evidencia), para obtener la probabilidad posterior de las demás variables

Inferencia en una RB

- Calcular la probabilidad a posteriori para un conjunto de variables de consulta, dado que se han observado algunos valores para las variables de evidencia. Por ejemplo, qué probabilidad hay de que realmente se haya producido un robo, sabiendo que tanto Juan como María han llamado a la policía P(Robo | J, M)
- Notación:

- X denotará la variable de consulta (sin pérdida de generalidad supondremos sólo una variable)

- E denota un conjunto de variables de evidencia E1, E2, . . . , En y e una observación concreta para esas variables

- Y denota al conjunto de las restantes variables de la red e y representa un conjunto cualquiera de valores para esas variables.

Inferencia en una RB

Definición formal

Supondremos que

- La RB está formada por las variables $\{X_1, X_2, \ldots, X_n\}$
- Las variables de interés son $X_I = \{X_1, X_2, \ldots, X_i\}$
- Las variables observadas (con evidencia) son: $X_O = \{X_{i+1}, \ldots, X_j\}$
  - Los valores que toman dichas variables (evidencia) son $e = \{e_{i+1}, \ldots, e_j\}$
- El resto de las variables es el conjunto $X_R = \{X_{j+1}, \ldots, X_n\}$

- El problema a resolver es calcular:

$$\boxed{\begin{array}{l} P(\mathbf{X}_I \mid \mathbf{X}_O = \mathbf{e}) = \dfrac{P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e})}{P(\mathbf{X}_O = \mathbf{e})} = \\[2ex] \dfrac{P(X_1, X_2, \ldots, X_i, X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j)}{P(X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j)} \end{array}}$$

Inferencia en una RB

El problema es calcular $$P(\mathbf{X}_I \mid \mathbf{X}_O = \mathbf{e}) = \frac{P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e})}{P(\mathbf{X}_O = \mathbf{e})}$$

El numerador es igual a

$$P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}) = \sum_{\mathbf{X}_R} P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}, \mathbf{X}_R) =$$

$$\sum_{X_{j+1}, X_{j+2}, \ldots, X_n} P(X_1, X_2, \ldots, X_i, X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j, X_{j+1}, X_{j+2}, \ldots, X_n)$$

El denominador es igual a

$$P(\mathbf{X}_O = \mathbf{e}) = \sum_{\mathbf{X}_I} P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}) =$$

$$\sum_{X_1, X_2, \ldots, X_i} P(X_1, X_2, \ldots, X_i, X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j)$$

Si se tiene el numerador, el denominador se puede calcular a partir de este

Algoritmos para calcular la Inferencia

– Algoritmos exactos

- Calcular de forma exacta la inferencia solicitada.
- La complejidad de resolver de forma exacta el problema general de inferencia en Redes Bayesianas es NP-duro.
- Tipos
  – Para redes específicas: Árboles (Pearl), (Complejidad lineal), Poliárboles (Kim, Pearl), (Complejidad lineal)
  – Para redes generales: **Eliminación de Variables**, Árbol de uniones (Lauritzen y Spiegelhalter)

– Algoritmos aproximados

- Se basan calcular de forma aproximada la inferencia solicitada simulando la distribución de la red bayesiana.
- Aproximar una distribución con una tolerancia dada es también NP-duro.
- Algunos algoritmos: Muestreo lógico (Henrion), Ponderación de la verosimilitud (Fung y Chang)

Algoritmo de Eliminación de Variables

$$P(\mathbf{X}_I \mid \mathbf{X}_O = \mathbf{e}) = \frac{P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e})}{P(\mathbf{X}_O = \mathbf{e})} =$$

$$\frac{P(X_1, X_2, \ldots, X_i, X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j)}{P(X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j)}$$

El problema es calcular

$$P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}) = \sum_{\mathbf{X}_R} P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}, \mathbf{X}_R) =$$

$$\sum_{X_{j+1}, X_{j+2}, \ldots, X_n} P(X_1, X_2, \ldots, X_i, X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j, X_{j+1}, X_{j+2}, \ldots, X_n)$$

Para realizar de forma eficiente la suma anterior:

1) Usar el teorema de factorización para factorizar la distribución conjunta

$$P(X_1, X_2, \ldots, X_i, X_{i+1}, X_{i+2}, \ldots, X_j, X_{j+1}, X_{j+2}, \ldots, X_n) = \prod P(X_i, pa(X_i))$$

2) Fijar las variables observadas a sus valores de evidencia

3) Eliminar de forma iterativa del sumatorio las variables que no son de interés ni de evidencia.

Algoritmo de Eliminación de Variables

Un **factor** se representa por una tabla para cada combinación de variables

Ejemplo f_X (*Temperatura, Tiempo*)

Temperatura	***Tiempo***	***P***
Calor	Sol	0.4
Calor	Lluvia	0.1
Frio	Sol	0.2
Frio	Lluvia	0.3

Tipos de **factores para probabilidades conjuntas** :

1) Distribución conjunta: P(A,B) -> f(A,B) - combinación en conjunto (toda la matriz)
2) Conjunción seleccionada: P(a,B) -> f(a,B) - es una **submatriz** de la anterior ya que a es un valor concreto. Por ejemplo, P(Calor, B)=0,4+0,1 o P(A, Sol) =0,4+0,2

Algoritmo de Eliminación de Variables

Tiempo = A	*Temperatura = B*	*P*
Sol	Calor	0.8
Lluvia	Calor	0.2
Sol	Frio	0.4
Lluvia	Frio	0.6

¡No es igual que antes!
Ahora son probabilidades condicionadas y no conjuntas

P(Tiempo | Calor)

P(Tiempo | Frio)

Tipos de **factores para probabilidades condicionadas**:

1) Condicional simple: P(A | b) -> f(A, b) – segunda variable fijada (la suma es 1). Por ejemplo, si fijamos Calor, entonces P(Sol|Calor)+P(Lluvia|Calor)=0.8+0.2=1
2) Familia de condicionales: P(A | B) -> f(A, B) - es una combinación de condicionales simples (la suma es el número de valores de la variable B (|B|)
3) Familia específica: P(a| B) -> f(a, B) – primera variable fijada (suma incierta). Por ejemplo, si fijamos Sol sumaría P(Sol | Calor)+P(Sol | Frio) = 0.8 + 0.4 = 1.2

Algoritmo de Eliminación de Variables

Multiplicación punto a punto o Unión de factores

La multiplicación de dos factores crea un nuevo factor cuyas variables son la unión de las variables y cuyos elementos son la multiplicación de los elementos iniciales

$$f_1(X_1,\ldots,X_i,Y_1,\ldots,Y_j) \cdot f_2(Y_1,\ldots,Y_j, Z_1,\ldots,Z_k) =$$

$$= f(X_1,\ldots,X_i,Y_1,\ldots,Y_j,Z_1,\ldots,Z_k)$$

Algoritmo de Eliminación de Variables

Multiplicación punto a punto o Unión de factores

$f_1(X_1,\ldots,X_i,Y_1,\ldots,Y_j) \,.\, f_2(Y_1,\ldots,Y_j, Z_1,\ldots,Z_k) = f(X_1,\ldots,X_i,Y_1,\ldots,Y_j,Z_1,\ldots,Z_k)$

Ejemplo

A	B	$f_1(A, B)$
T	T	0.3
T	F	0.7
F	T	0.9
F	F	0.1

B	C	$f_2(B, C)$
T	T	0.2
T	F	0.8
F	T	0.6
F	F	0.4

A	B	C	f(A, B, C)
T	T	T	0.3x0.2=0.06
T	T	F	0.3x0.8=0.24
T	F	T	0.7x0.6=0.42
T	F	F	0.7x0.4=0.28
F	T	T	0.9x0.2=0.18
F	T	F	0.9x0.8=0.72
F	F	T	0.1x0.6=0.06
F	F	F	0.1x0.4=0.04

Algoritmo de Eliminación de Variables

Multiplicación punto a punto o Unión de factores

Con variables binarias, si el primer factor tiene 2^{i+j} elementos y el segundo 2^{j+k} elementos, entonces el factor resultante tiene 2^{i+j+k} elementos

La **complejidad es exponencial** tanto en tiempo como en espacio

Algoritmo de Eliminación de Variables

Suma de factores (eliminación)

Dentro de cada factor, se suman las submatrices que se obtienen al fijar cada uno de los valores posibles de las variables que no se van a eliminar

$$\mathbf{f(A, B, C,\ldots,N,E) = f(A, B, C,\ldots,N)}$$

E = variables a eliminar

Algoritmo de Eliminación de Variables

Suma de factores (eliminación)

f(A, B, C) = f(A, C)

A	B	C	f(A,B,C)	A	C	f(A,C)
T	T	T	0.3x0.2=0.06	T	T	0.06 + 0.42 = 0.48
T	T	F	0.3x0.8=0.24	T	F	0.24 + 0.28 = 0.52
T	F	T	0.7x0.6=0.42	F	T	0.18 + 0.06 =0.24
T	F	F	0.7x0.4=0.28	F	F	0.72 + 0.04 = 0.76
F	T	T	0.9x0.2=0.18			
F	T	F	0.9x0.8=0.72			
F	F	T	0.1x0.6=0.06			
F	F	F	0.1x0.4=0.04			

Algoritmo de Eliminación de Variables

Orden de eliminación de las variables

Para eliminar las variables ocultas se utilizan los dos pasos anteriores

La mejor estrategia es la **voraz**: la siguiente variable a eliminar será la que minimice el tamaño del siguiente factor a crear

Para ello, es mejor empezar por los factores que contienen las **evidencias**
f(A,C); f(B,C); f(A, B, e)

Algoritmo de Eliminación de Variables

RESUMEN

El Algoritmo de eliminación de variables evita la repetición de cálculos realizándolos una sola vez y guardándolos para su reutilización

A los valores guardado se les llama **factores**

Los primeros factores son las probabilidades condicionadas iniciales, después aparecen nuevos factores que son las multiplicaciones entre ellos almacenados en matrices de probabilidad

El orden en que se eliminan las variables influye mucho en la complejidad de la red

Un factor corresponde a la probabilidad de un conjunto de variables dadas otras variables

Algoritmo de Eliminación de Variables

ALGORITMO

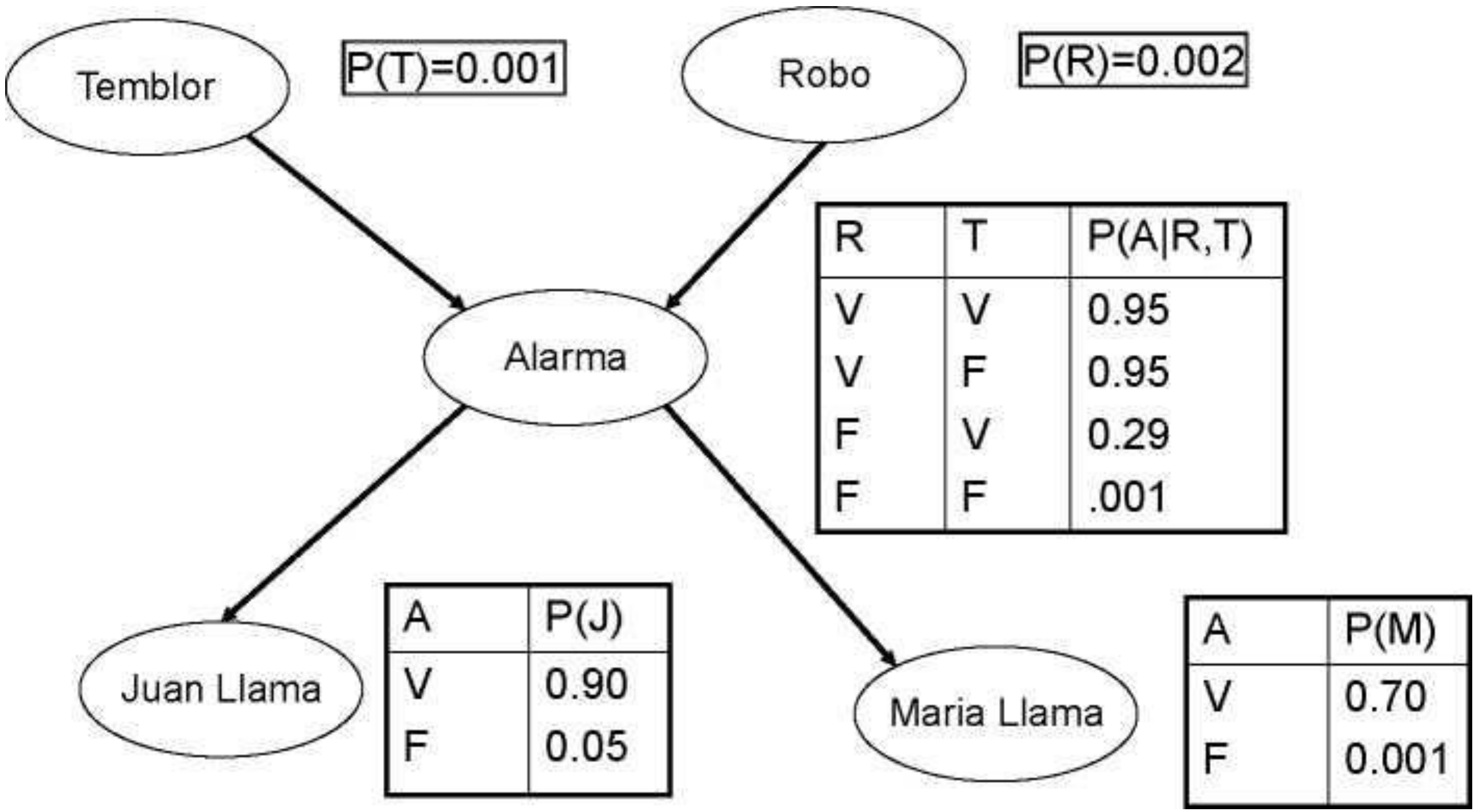
Se ejecuta de forma iterativa los siguientes **pasos**:

1) Se elige una variable oculta a eliminar
2) Se seleccionan los factores donde interviene esa variable
3) Unión: multiplicación punto a punto
4) Suma: se elimina esa variable
5) Se eliminan todos los factores donde aparecía la variable eliminada
6) Si aún quedan variables ocultas, se repite el proceso
7) Normalización

Veamos un ejemplo de aplicación:

Algoritmo de Eliminación de Variables

Ejemplo



R	T	P(A|R,T)
V	V	0.95
V	F	0.95
F	V	0.29
F	F	.001

A	P(J)
V	0.90
F	0.05

A	P(M)
V	0.70
F	0.001

Algoritmo de Eliminación de Variables

Queremos calcular $P(R \mid j, m) \propto P(R, j, m)$ --> $f(R, j, m)$.

Con R la **variable pregunta** (mayúsculas); **j, m** son las **evidencias** (minúsculas) y las **variables ocultas** son T y A.

Según el teorema de factorización $$P(x_1,..,x_n) = \prod_{i=1..n} P(x_i \mid Padres(x_i))$$

$P(R, j, m) = P(R).P(T).P(A \mid R,T).P(j \mid A).P(m \mid A)$

Es decir, hay que calcular los factores correspondientes, es decir, $f_1(R)$, $f_2(T)$, $f_3(A, R, T)$, $f_4(A, j) = f_4(j, A)$ y $f_5(m, A) = f_5(A, m)$

Algoritmo de Eliminación de Variables

PASOS del ALGORITMO

1) Se elige una variable oculta a eliminar: A
2) Se seleccionan los factores donde interviene esa variable:

   $f_3(A, R, T)$, $f_4(j, A)$ y $f_5(A, m)$

3) Unión: multiplicación punto a punto

   $$f'_6(A, R, T, j, m) = f_3(A, R, T) \times f_4(j, A) \times f_5(A, m)$$

4) Suma: se elimina esa variable

   $f'_6(A, R, T, j, m) \rightarrow f_6(R, T, j, m)$

Algoritmo de Eliminación de Variables

PASOS del ALGORITMO

5) Se eliminan todos los factores donde aparecía la variable eliminada A

$$f_1(R),\ f_2(T),\ \cancel{f_3(A, R, T)},\ \cancel{f_4(j, A)},\ \cancel{f_5(A, m)},\ f_6(R, T, j, m)$$

6) Si aún quedan variables ocultas, se repite el proceso: eliminar T

$$f_1(R),\ \cancel{f_2(T)},\ \cancel{f_6(R, T, j, m)},\ f_7(R, j, m)$$

7) Realizar el producto y normalizar

$$P(R \mid j,m) \leftarrow \alpha\ f_1(R) \times f_7(R, j, m)$$

Algoritmo de Eliminación de Variables

```
función ELIMINACIÓN-VARIABLES(preguntas,
                              evidencias,
                              red-bayesiana)
devuelve distribución
  factores ← CREAR-LISTA(red-bayesiana.TABLAS)
  variables ← red-bayesiana.VARIABLES
  ocultas ← red-bayesiana.OCULTAS.QUITAR(preguntas)
  por cada variable en ORDENAR(variables) hacer
    factor ← CREAR-FACTOR(variable, evidencias)
    factores.AGREGAR(factor)
    si variable en ocultas entonces
      suma ← SUMAR(variables, factores)
      factores.AGREGAR(suma)
  producto ← MULTIPLICAR(factores)
devolver NORMALIZAR(producto)
```

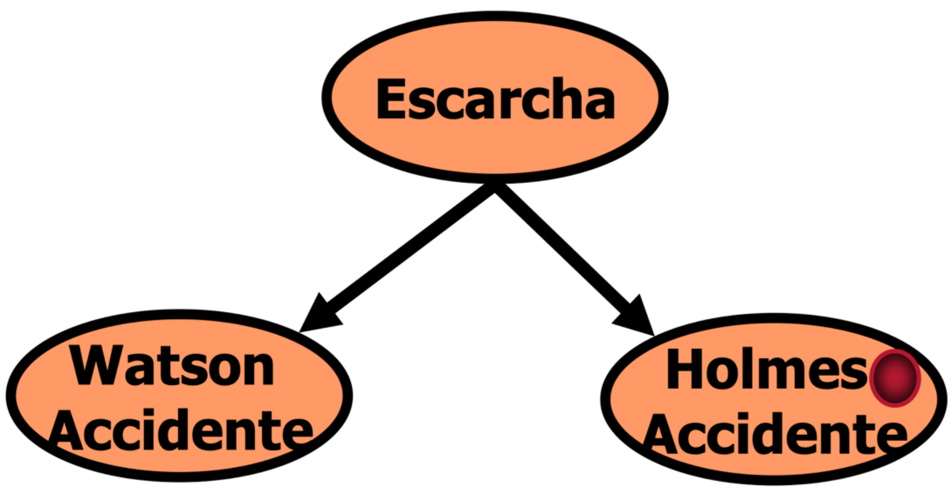
Algoritmo de Eliminación de Variables

Ejemplo. Calcular P(W | h) - con evidencias

E	0.7
¬E	0.1

	W	¬W
E	0.8	0.2
¬E	0.1	0.9

	H	¬H
E	0.8	0.2
¬E	0.1	0.9



Algoritmo de Eliminación de Variables

Ejemplo: calcular $P(W \mid h) \propto P(W, h)$

Preparación

Variable pregunta W (mayúsculas)

Evidencias h (minúsculas)

Variables ocultas E (variables a eliminar)

Factorización

$P(W \mid H) = P(E).P(W \mid E).P(H \mid E)$

Sustituir evidencias

$P(E).P(W \mid E).P(h \mid E)$

$f_1(E).\ f_2(W, E).f_3(h, E)$

Algoritmo de Eliminación de Variables

PASOS

1) Se elige una variable oculta a eliminar: E
2) Se seleccionan los factores donde interviene esa variable:

   $f_1(E)$, $f_2(W, E).f_3(h, E)$

3) Unión: multiplicación punto a punto

   $f'_4(W, E, h) = f_1(E).f_2(W, E).f_3(h, E)$

4) Suma: se elimina esa variable (h es una evidencia, no varía)

   1) $f'_4(W, E, h) = \sum_E f_1(E).f_2(W, E).f_3(h, E) \rightarrow f_4(W, h) = 0.7 \times 0.8 \times 0.8 + 0.3 \times 0.1 \times 0.1 = 0.451$

   2) $f'_4(W, E, h) = \sum_E f_1(E).f_2(W, E).f_3(h, E) \rightarrow f_4(\neg W, h) = 0.7 \times 0.2 \times 0.8 + 0.3 \times 0.9 \times 0.1 = 0.139$

5. Se eliminan todos los factores donde aparecía la variable E

   $f_4(W, h) \leftarrow P(W \mid h) \rightarrow f_4(\neg W, h)$

6. Normalizamos → 0.451 + 0.139 = 0.59 luego las normalizaciones son 0.451 / 0.59 = 0.76 para W y 0.139 / 0.59 = 0.24

Algoritmo de Eliminación de Variables

Ejemplo: calcular P(W)

Preparación

Variable pregunta W (mayúsculas)

Evidencias ninguna (minúsculas)

Variables ocultas H, E (variables a eliminar)

Factorización

P(W) = P(E).P(W | E).P(H | E)

Sustituir evidencias

P(E).P(W | E).P(H | E)

$f_1(E).\ f_2(W, E).f_3(H, E)$

Algoritmo de Eliminación de Variables

PASOS

1) Eliminamos la variable oculta H
2) Se seleccionan los factores donde interviene esa variable:

   $f_3(H, E)$

3) Unión: multiplicación punto a punto

   $f_3(H, E)$

4) Suma: se elimina esa variable H

   $f'_4(H, E) \rightarrow \sum_H f_3(H, E) = \sum_H P(H \mid E) \rightarrow f_4(E) = 0.8 + 0.2 = 1$

   $f'_4(H, E) \rightarrow \sum_H f_3(H, \neg E) = \sum_H P(H \mid \neg E) \rightarrow f_4(\neg E)\ \ 0.1 + 0.9 = 1$

Algoritmo de Eliminación de Variables

PASOS

5. Eliminamos E
6. Unión: multiplicación punto a punto

$f'_5(E,H,W) = f_1(E).\ f_2(W, E).f_3(H, E)$

7. Suma: se elimina E

$f'_5(E, W) = \sum_E f_1(E)f_2(W, E)f_4(H) \rightarrow f_5(W) = 1 \times 0.7 \times 0.8 + 1 \times 0.3 \times 0.1 = 0.59$

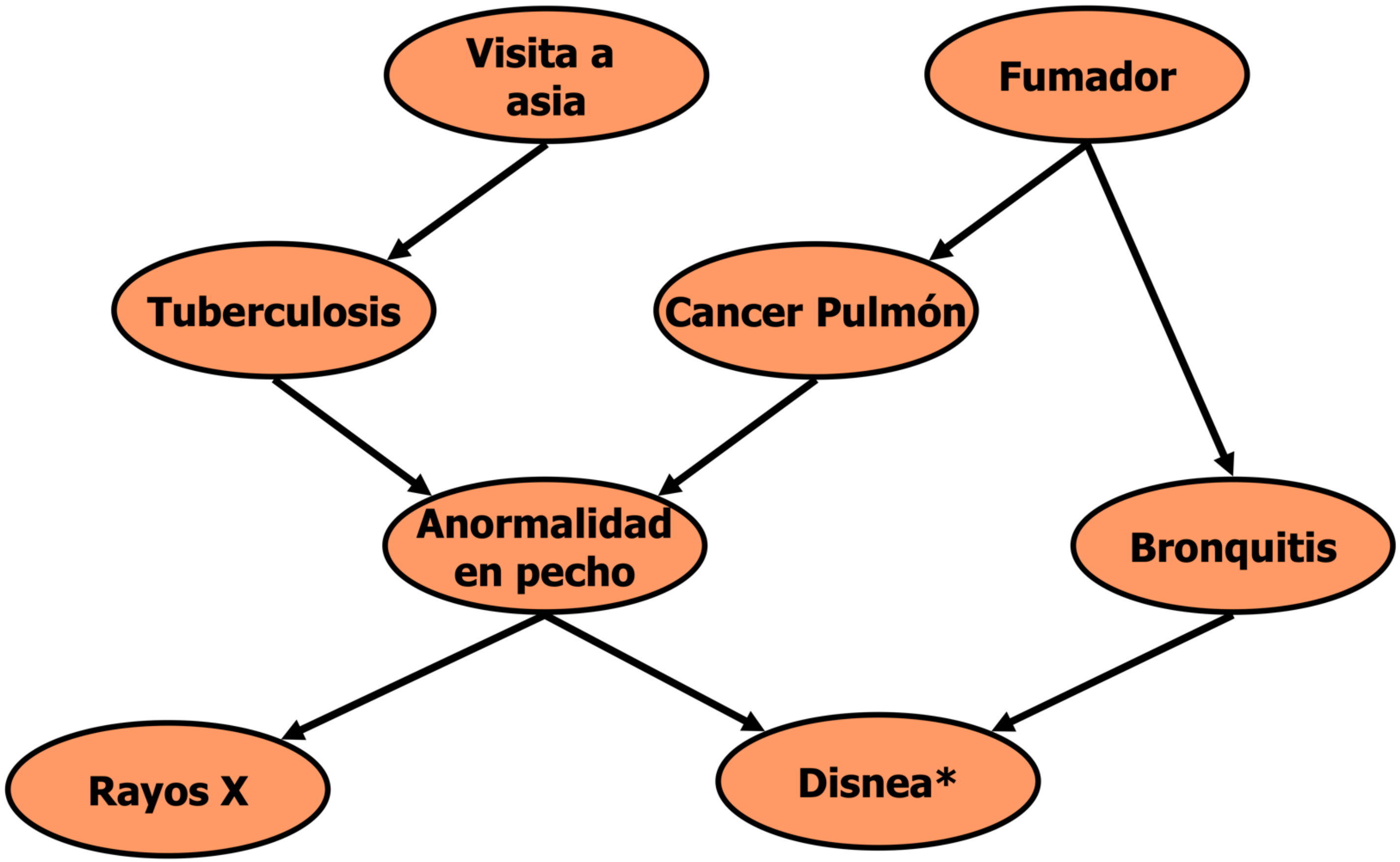
$f'_5(E, W) = \sum_E f_1(E)f_2(\neg W, E)f_4(H) \rightarrow f_5(\neg W) = 1 \times 0.7 \times 0.2 + 1 \times 0.3 \times 0.9 =$
0.41

8. Normalizamos → Ya está normalizado

Algoritmo de Eliminación de Variables

Ejemplo

La Red Asia



*Disnea=Dificultad para respirar

Algoritmo de Eliminación de Variables

Ejemplo: calcular $P(C \mid v, f, d) \propto P(C, v, f, d)$

Preparación

Variable pregunta C (mayúsculas)

Evidencias v, f, d (minúsculas)

Variables ocultas T, A, B, R (variables a eliminar)

Factorización

$P(C \mid v, f, d) = P(V).P(F).P(T \mid V).P(C \mid F).\ P(B \mid F).\ P(A \mid T, C).\ P(R \mid A).\ P(D \mid A, B)$

Sustituir evidencias

$P(v).P(f).P(T \mid v).P(C \mid f).\ P(B \mid f).\ P(A \mid T, C).\ P(R \mid A).\ P(d \mid A, B)$

$f_1(v).f_2(f).f_3(T, v).f_4(C, f).\ f_5(B, f).\ f_6(A, T, C).\ f_7(R, A).\ f_8(d, A, B)$

Algoritmo de Eliminación de Variables

PASOS

1) Se elige una variable oculta a eliminar: R
2) Se seleccionan los factores donde interviene esa variable:

   $f_7(R, A)$

3) Unión: multiplicación punto a punto

   $f'_9(R, A)$

4) Suma: se elimina esa variable

   $f'_9(R, A) \rightarrow f_9(A)$

5. Se eliminan todos los factores donde aparecía la variable R

   $f_1(v).f_2(f).f_3(T, v).f_4(C, f).\ f_5(B, f).\ f_6(A, T, C).\ f_8(d, A, B).\ f_9(A)$

6. Si aún quedan variables ocultas, se repite el proceso: ahora eliminamos T

   $f_1(v).f_2(f).f_4(C, f).\ f_5(B, f).\ f_8(d, A, B).\ f_9(A).\ f_{10}(A, C)$

Algoritmo de Eliminación de Variables

PASOS

7. Si aún quedan variables ocultas, se repite el proceso: eliminamos A

   $f_1(v).f_2(f).f_4(C, f). f_5(B, f). f_{11}(C, B)$

8. Si aún quedan variables ocultas, se repite el proceso: eliminamos B

   $f_1(v).f_2(f).f_4(C, f). f_{12}(C)$

9. Como no quedan variables ocultas, se multiplica y normaliza

Algoritmo de Eliminación de Variables

Ejemplo: calcular P(D)

Preparación

Variable pregunta D (mayúsculas)

Evidencias ninguna

Variables ocultas V, F, R, T, C, A, B (variables a eliminar)

Factorización

P(D) = P(V).P(F).P(T | V).P(C | F). P(B | F). P(A | T, C). P(R | A). P(D | A, B)

Sustituir evidencias

P(V).P(F).P(T | V).P(C | F). P(B | F). P(A | T, C). P(R | A). P(D | A, B)

f_1(V).f_2(F).f_3(T, V).f_4(C, F). f_5(B, F). f_6(A, T, C). f_7(R, A). f_8(D, A, B)

Algoritmo de Eliminación de Variables

PASOS

Eliminamos V

$f_2(F).f_4(C, F).\ f_5(B, F).\ f_6(A, T, C).\ f_7(R, A).\ f_8(D, A, B).f_9(T)$

Eliminamos F

$f_6(A, T, C).\ f_7(R, A).\ f_8(D, A, B).f_9(T).f_{10}(C, B)$

Eliminamos R

$f_6(A, T, C).\ f_8(D, A, B).f_9(T).f_{10}(C, B).f_{11}(A)$

Eliminamos T

$f_8(D, A, B).\ f_{10}(C, B).f_{11}(A).f_{12}(A, C)$

Eliminamos C

$f_8(D, A, B).\ f_{11}(A).f_{13}(A, B)$

Eliminamos A

$f_{14}(D, B)$

Eliminamos B

$f_{15}(D)$

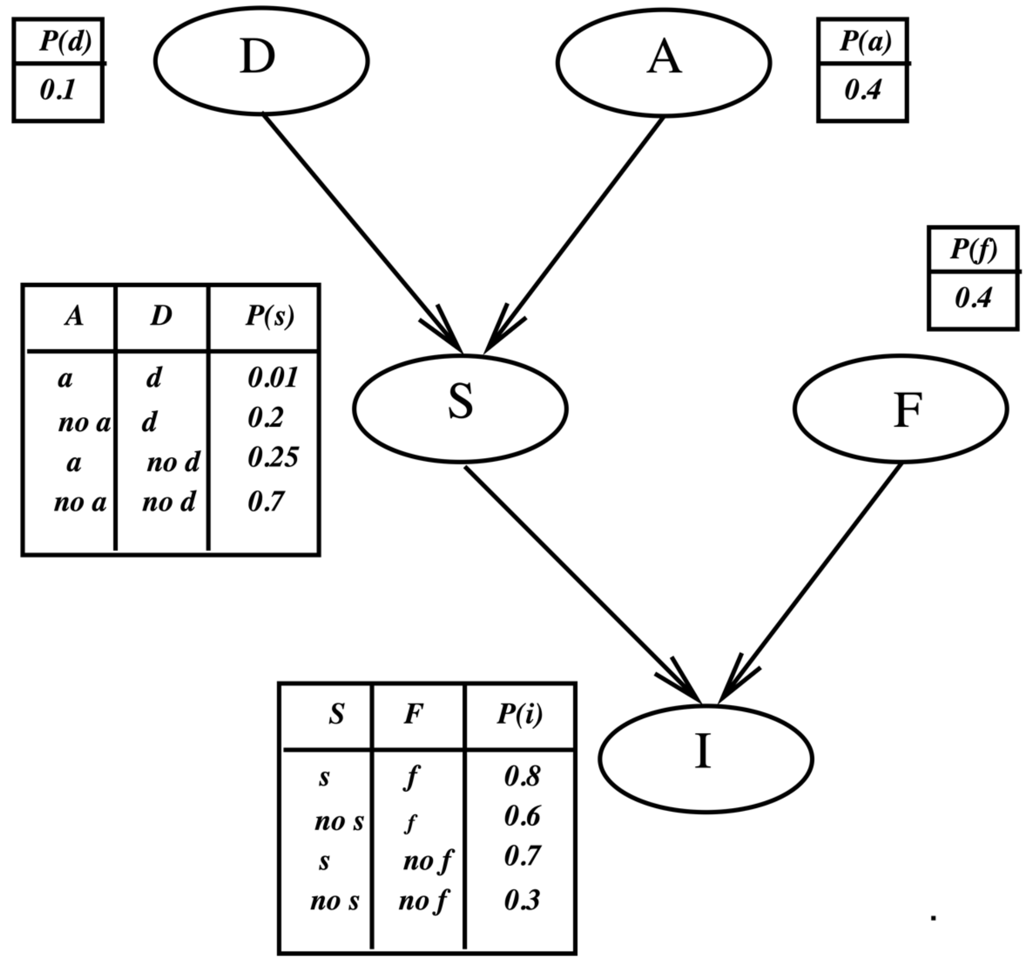
Ejercicio

Consideremos las siguientes variables aleatorias: D -> práctica deportiva habitual; A -> alimentación equilibrada, S -> presión sanguínea alta; F -> fumador; I -> ha sufrido un infarto de miocardio.

Las relaciones causales y el conocimiento probabilístico asociado están reflejadas en la siguiente red bayesiana.

Calcula la probabilidad de ser fumador si se ha sufrido un infarto y no se hace deporte, es decir

$$P(F\ |i,\neg d) = \propto = P(F,i,\neg d)$$



P(d)
0.1

P(a)
0.4

P(f)
0.4

A	D	P(s)
a	d	0.01
no a	d	0.2
a	no d	0.25
no a	no d	0.7

S	F	P(i)
s	f	0.8
no s	f	0.6
s	no f	0.7
no s	no f	0.3

Ejercicio

Dada la red de la figura, calcula las probabilidades marginales y condicionales a) Pr(¬p3), b) Pr(p2|¬p3), c) Pr(p1|p2,¬p3) y d) Pr(p1|¬p3,p4).

Soluciones:
a) 0.762
b) 0.6509
c) 0.5161
d) 0.4394