Tema 7: Razonamiento con incertidumbre

J. L. Ruiz Reina
F.J. Martín Mateos
M.A. Gutiérrez Naranjo

Dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla

Contenido

- Incertidumbre y conocimiento. Variables aleatorias
- Probabilidad incondicional
- Probabilidad condicional: inferencia probabilística
- Independencia
- Independencia condicional
- Redes bayesianas
- Inferencia exacta en redes bayesianas
- Inferencia aproximada en redes bayesianas
- Naive Bayes

Representación del conocimiento

Algunos formalismos de representación:

- Reglas, lógicas de descripción, lógica de primer orden, redes bayesianas . . .
- Cada formalismo de representación usa un método de inferencia específico:
  - Razonamiento hacia adelante, razonamiento hacia atrás (SLD-resolución), tableros, cálculo de probabilidades . . .

En este tema usaremos la teoría y cálculo de probabilidades para representar el conocimiento y razonar a partir de él

Incertidumbre y conocimiento

- Ejemplo de conocimiento "categórico" expresado mediante reglas:

  Si Sintoma=dolor de muelas
  Entonces Enfermedad=caries

- ¿Expresa esta regla un conocimiento correcto?
- Quizás sería mejor un conocimiento más exhaustivo

  Si Sintoma=dolor de muelas
  Entonces Enfermedad=caries O
  Enfermedad=sinusitis O
  Enfermedad=muela del juicio O
  ...

Inconvenientes del conocimiento categórico

- ¿Por qué a veces no podemos tener conocimiento exacto y preciso?
- Posibles razones
  - Demasiado trabajo ser exhaustivo
  - Desconocimiento teórico: no tenemos información completa
  - Desconocimiento práctico: aún conociendo todas las reglas, no podemos aplicarlas
  - No determinismo

Incertidumbre y conocimiento

- Otra forma de expresar el conocimiento: grado de creencia
  - *Creemos*, basándonos en nuestras *percepciones*, que un paciente que tenga dolor de muelas, tiene caries con una probabilidad del 80 %
  - En teoría de la probabilidad, se expresa como $P(Caries = true | Dolor = true) = 0{,}8$
  - La probabilidad expresa el *grado de creencia*, no el *grado de verdad*
  - Por tanto, la probabilidad puede cambiar a medida que se conocen nuevas *evidencias*
- La *teoría de la probabilidad* servirá como medio de representación del conocimiento incierto

Variables aleatorias

- Variables aleatorias: una "parte" del mundo cuyo estado podemos desconocer
  - Ejemplo: la variable aleatoria *Caries* describe el hecho de que un paciente pueda o no tener caries
  - Nuestra descripción del mundo vendrá dada por un conjunto de variables aleatorias
- Una variable aleatoria puede tomar diferentes *valores* de su *dominio*
  - Los posibles valores de *Caries* son *true* y *false*
  - Notación: variables aleatorias en mayúsculas y sus valores en minúsculas

Variables aleatorias

- Tipos de variables aleatorias:
  - Booleanas (notación: *caries* y $\neg$*caries* son equivalentes a *Caries* = *true* y *Caries* = *false*, respectivamente)
  - Discretas (incluyen a las booleanas)
  - Continuas
- En lo que sigue, nos centraremos en las variables discretas

Proposiciones

- Usando las conectivas proposicionales y las variables aleatorias, podemos expresar *proposiciones*
- Ejemplos:
  - $caries \wedge \neg dolor$
  - $Caries = true \vee Tiempo = nublado$
  - $Tirada1 = 5 \wedge (Tirada2 = 4 \vee Tirada2 = 5 \vee Tirada2 = 6)$
- Asignaremos probabilidades a las proposiciones para expresar nuestro grado de creencia en las mismas

Probabilidad incondicional: idea intuitiva

- Dada una proposición a, su probabilidad incondicional (o *a priori*), notada $P(a)$, cuantifica el grado de creencia en que ocurra a, *en ausencia de cualquier otra información o evidencia*
  - Ejemplo: $P(caries) = 0{,}1$, $P(caries, \neg dolor) = 0{,}05$
  - Notación: $P(caries, \neg dolor)$ es equivalente a $P(caries \wedge \neg dolor)$
- Aproximación *frecuentista*: número de casos favorables (en los que se cumple a) entre el número de casos totales

Probabilidad: definición axiomática

- Una función de probabilidad es una función definida en el conjunto de proposiciones (respecto de un conjunto dado de variables aleatorias), verificando las siguientes propiedades:
  - $0 \leq P(a) \leq 1$, para toda proposición a
  - $P(true) = 1$ y $P(false) = 0$ donde *true* y *false* representan a cualquier proposición tautológica o insatisfactible, respectivamente
  - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$, para cualquier par de proposiciones a y b
- El *cálculo de probabilidades* se construye sobre los tres axiomas anteriores. Por ejemplo:
  - $P(\neg a) = 1 - P(a)$
  - $P(a \vee b) = P(a) + P(b)$, si a y b son disjuntos.
  - $\sum_{i=1}^{n} P(D = d_i) = 1$, siendo D una v.a. y $d_i, i = 1, \ldots, n$ sus posibles valores

Distribuciones de probabilidad

- La *distribución de probabilidad* de una variable aleatoria indica las probabilidades de que la variable pueda tomar cada uno de sus valores
  - Ejemplo: si *Tiempo* es una v.a. con valores *lluvia*, *sol*, *nubes* y *nieve*, su distribución de probabilidad podría ser:
  - $P(\mathit{Tiempo} = \mathit{sol}) = 0{,}7$, $P(\mathit{Tiempo} = \mathit{lluvia}) = 0{,}2$, $P(\mathit{Tiempo} = \mathit{nubes}) = 0{,}08$, $P(\mathit{Tiempo} = \mathit{nieve}) = 0{,}02$
- Notación: usaremos **P** (en negrita), para expresar de manera compacta una distribución de probabilidad (fijado un orden entre sus valores)
  - Ejemplo: $\mathbf{P}(\mathit{Tiempo}) = \langle 0{,}7; 0{,}2; 0{,}08; 0{,}02 \rangle$

Distribución conjunta

- Distribución de probabilidad *conjunta*: probabilidad de cada combinación de valores de dos o más variables aleatorias
  - Notación $\mathbf{P}(X, Y)$: manera compacta de denotar a una tabla con esas probabilidades
  - Ejemplo: $\mathbf{P}(\mathit{Tiempo}, \mathit{Caries})$ denota una tabla con 4×2 entradas

Eventos atómicos

- Dado un conjunto de variables aleatorias que describen nuestro "mundo", un *evento atómico* es un tipo particular de proposición:
  - Conjunción de proposiciones elementales, que expresan un valor concreto para todas y cada una de las variables
- Ejemplo: si *Caries* y *Dolor* son todas las variables aleatorias en nuestra descripción del mundo, los posibles eventos atómicos son
  - $caries \wedge dolor$
  - $caries \wedge \neg dolor$
  - $\neg caries \wedge dolor$
  - $\neg caries \wedge \neg dolor$

Eventos atómicos

- Características de los eventos atómicos:
  - Mútuamente excluyentes
  - Todos los eventos atómicos son exhaustivos (alguno debe ocurrir)
  - Un evento atómico implica la verdad o falsedad de toda proposición
  - Toda proposición es equivalente a la disyunción de un conjunto de eventos atómicos: por ejemplo, *caries* es equivalente a $(caries \wedge dolor) \vee (caries \wedge \neg dolor)$
  - Para cualquier proposición a,

$$P(a) = \sum_{e_i \in \mathbf{e}(a)} P(e_i)$$

siendo $\mathbf{e}(a)$ el conjunto de eventos atómicos cuya disyunción equivale a a

Distribución conjunta y completa

- Distribución de probabilidad *conjunta y completa* (DCC): probabilidad de cada evento atómico
  - Una DCC es una *especificación completa* (en términos probabilísticos) del dominio descrito
- Ejemplo de DCC:

	$dolor$	$dolor$	$\neg dolor$	$\neg dolor$
	$hueco$	$\neg hueco$	$hueco$	$\neg hueco$
$caries$	0.108	0.012	0.072	0.008
$\neg caries$	0.016	0.064	0.144	0.576

Cálculo de probabilidades usando DCC

- Usando la fórmula $P(a) = \sum_{e_i \in \mathbf{e}(a)} P(e_i)$
- Ejemplos:
  - $P(caries \vee dolor) =$
    $P(caries, dolor, hueco) + P(caries, dolor, \neg hueco) +$
    $P(caries, \neg dolor, hueco) + P(caries, \neg dolor, \neg hueco) +$
    $P(\neg caries, dolor, hueco) + P(\neg caries, dolor, \neg hueco) =$
    $0{,}108 + 0{,}012 + 0{,}072 + 0{,}008 + 0{,}016 + 0{,}064 = 0{,}28$
  - $P(caries) =$
    $P(caries, dolor, hueco) + P(caries, dolor, \neg hueco) +$
    $P(caries, \neg dolor, hueco) + P(caries, \neg dolor, \neg hueco) =$
    $0{,}108 + 0{,}012 + 0{,}072 + 0{,}008 = 0{,}2$

Cálculo de probabilidades usando DCC

- En general:
  - $\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{Z}} \mathbf{P}(\mathbf{Y}, \mathbf{z})$ (regla de *marginalización*)
- Notación
  - $\mathbf{Y}$ es un vector de variables aleatorias, simbolizando cualquier combinación de valores de esas variables
  - $\mathbf{z}$ representa una combinación de valores concretos para un conjunto $\mathbf{Z}$ de variables aleatorias (las restantes)
  - Hay un sumando $\mathbf{P}(\mathbf{Y}, \mathbf{z})$ para cada posible $\mathbf{z}$, y cada sumando es una entrada de la DCC
- Problemas:
  - Tamaño exponencial de la DCC
  - Rara vez se conocen directamente las probabilidades de todos los eventos atómicos.
- Las *probabilidades condicionales* expresan mejor nuestro conocimiento del dominio.

Probabilidad condicional

- Probabilidad *condicional* (o *a posteriori*) asociada a a dado b (a y b proposiciones):
  - Grado de creencia sobre a, dado que *todo lo que sabemos* es que b ocurre, notada $P(a|b)$
  - Ejemplo: $P(caries|dolor) = 0{,}8$ significa que una vez sabido que un paciente tiene dolor de muelas (*y sólamente sabemos eso*), nuestra creencia es que el paciente tendrá caries con probabilidad 0,8

Probabilidad condicional

- Relación entre probabilidad condicional e incondicional:

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

- Variante: $P(a \wedge b) = P(a|b)P(b)$ (*regla del producto*)
  - O, análogamente, $P(a \wedge b) = P(b|a)P(a)$
- Notación $\mathbf{P}(X|Y)$ para expresar la tabla de probabilidades condicionales
  - Forma compacta de la regla del producto: $\mathbf{P}(X, Y) = \mathbf{P}(X|Y)\mathbf{P}(Y)$
  - Nota: la fórmula anterior *no expresa* un producto de matrices, sino la relación existente entre las correspondientes entradas de las tablas

Probabilidad condicional *vs* implicación lógica

- La probabilidad condicional formaliza el hecho de que los grados de creencia se actualizan a medida que se van conociendo nuevas evidencias en el mundo incierto
- La probabilidad condicional *no* es lo mismo que una implicación lógica con incertidumbre
  - $P(a|b) = 0{,}8$ no es lo mismo que decir que "siempre que b sea verdad, entonces $P(a) = 0{,}8$"
  - Ya que $P(a|b)$ refleja que b es *la única* evidencia conocida

Inferencia probabilística

- Por *inferencia probabilística* entendemos el cálculo de la probabilidad de una proposición dada *condicionada* por la observación de determinadas evidencias
  - Es decir, cálculos del tipo $P(a|b)$ donde a es la proposición que se *consulta* y b es la proposición que se ha *observado*
  - El *conocimiento base* vendrá dado por una DCC (representada de alguna manera *eficiente*, a partir de otras probabilidades condicionales ya conocidas, como ya veremos)
- El estudio de algoritmos de inferencia probabilística es el principal objetivo de este tema

Inferencia probabilística a partir de una DCC

- En principio, podemos calcular *probabilidades condicionales* usando la DCC
- Por ejemplo: probabilidad de tener *caries*, observado que hay *dolor*

$$P(caries|dolor) = \frac{P(caries \wedge dolor)}{P(dolor)} =$$

$$= \frac{0{,}108 + 0{,}012}{0{,}108 + 0{,}012 + 0{,}016 + 0{,}064} = 0{,}6$$

Normalización

- Podríamos evitar calcular explícitamente $P(dolor)$
- $\frac{1}{P(dolor)}$ puede verse como una constante que *normaliza* la distribución $\mathbf{P}(Caries, dolor)$ haciendo que sume 1:

$$\mathbf{P}(Caries|dolor) = \alpha\mathbf{P}(Caries, dolor) =$$

$$= \alpha[\mathbf{P}(Caries, dolor, hueco) + \mathbf{P}(Caries, dolor, \neg hueco)] =$$

$$= \alpha[\langle 0{,}108; 0{,}016\rangle + \langle 0{,}012; 0{,}064\rangle] = \alpha\langle 0{,}12; 0{,}08\rangle = \langle 0{,}6; 0{,}4\rangle$$

- Es decir, calculamos $P(caries, dolor)$ y $P(\neg caries, dolor)$ y a posteriori multiplicamos ambos por una constante α que haga que ambos sumen 1; de esa manera tenemos $P(caries|dolor)$ y $P(\neg caries|dolor)$

Inferencia probabilística a partir de una DCC

- En general, dada una variable aleatoria X, un conjunto de variables observadas $\mathbf{E}$ (con valor concreto $\mathbf{e}$), se tiene:

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

  - es una igualdad entre "vectores" (una componente por cada posible valor de X)
  - hay un sumando por cada combinación $\mathbf{y}$ de valores de variables $\mathbf{Y}$ no observadas
  - para cada valor de X, cada sumando $\mathbf{P}(X,\mathbf{e},\mathbf{y})$ es un elemento de la DCC
  - α es una constante de normalización, que hace que la distribución de probabilidades sume 1

Inferencia probabilística a partir de una DCC

- Dada una DCC, la fórmula anterior nos da un método para realizar inferencia probabilística
- Problema en la práctica: exponencialidad
  - Con n variables, procesar la DCC necesita un tiempo $O(2^n)$
  - En un problema real, podría haber cientos o miles de variables
- Sin embargo, veremos que la posible independencia existente entre los eventos que describen las distintas variables aleatorias nos puede ayudar a reducir esta complejidad

Independencia probabilística

- En muchos casos prácticos, algunas de las variables de un problema son *independientes* entre sí
  - Ejemplo: $P(Tiempo = nublado|dolor, caries, hueco) = P(Tiempo = nublado)$
  - Si la variable *Tiempo* (con 4 posibles valores) formara parte de una descripción en la que están *Caries*, *Hueco* y *Dolor*:
    - No necesitaríamos una tabla con 32 entradas para describir la DCC, sino dos tablas independientes (8+4 entradas)
    - Esto reduce la complejidad de la representación

Independencia probabilística

- Intuitivamente, dos variables aleatorias son independientes si conocer el valor que toma una de ellas no nos actualiza (ni al alza ni a la baja) nuestro grado de creencia sobre el valor que tome la otra.
  - El asumir que dos variables son independientes está basado normalmente en el conocimiento previo del dominio que se modela
- Formalmente, dos variables aleatorias X e Y son *independientes* si $\mathbf{P}(X|Y) = \mathbf{P}(X)$ (equivalentemente, $\mathbf{P}(Y|X) = \mathbf{P}(Y)$ ó $\mathbf{P}(X, Y) = \mathbf{P}(X) \cdot \mathbf{P}(Y)$)
  - En general, dos proposiciones a y b son independientes si $P(a|b) = P(a)$
- Asumir independencia entre ciertas variables ayuda a que la representación del mundo sea más manejable
  - Reduce la exponencialidad (*factorización* del problema)

Independencia condicional

- Sin embargo, en nuestro ejemplo *Dolor* y *Hueco no son independientes*
  - Ambas dependen de *Caries*
- Pero son independientes *una vez conocido* el valor de *Caries*
  - Es decir: $\mathbf{P}(Dolor|Hueco, Caries) = \mathbf{P}(Dolor|Caries)$ o equivalentemente $\mathbf{P}(Hueco|Dolor, Caries) = \mathbf{P}(Hueco|Caries)$
  - También equivalente: $\mathbf{P}(Hueco, Dolor|Caries) = \mathbf{P}(Hueco|Caries)\mathbf{P}(Dolor|Caries)$

Independencia condicional

- Intuitivamente: X es condicionalmente independiente de Y dado un conjunto de variables $\mathcal{Z}$ si nuestro grado de creencia en que X tome un valor dado, sabiendo el valor que toman las variables de $\mathcal{Z}$, no se vería actualizado (ni al alza ni a la baja), si además supiéramos el valor que toma Y
  - En el ejemplo: sabiendo si se tiene *Caries* o no, el conocer si se tiene *Dolor* o no, no va a aportar nada nuevo a nuestro grado de creencia en que se tenga *Hueco*.
- Formalmente, dos v.a. X e Y son independientes dado un conjunto de v.a. $\mathcal{Z}$ si $\mathbf{P}(X, Y|\mathcal{Z}) = \mathbf{P}(X|\mathcal{Z})\mathbf{P}(Y|\mathcal{Z})$
  - O equivalentemente $\mathbf{P}(X|Y, \mathcal{Z}) = \mathbf{P}(X|\mathcal{Z})$
    ó $\mathbf{P}(Y|X, \mathcal{Z}) = \mathbf{P}(Y|\mathcal{Z})$

Independencia condicional

- La independencia condicional entre algunas variables es esencial para un almacenamiento eficiente de las DCCs. Por ejemplo

$$\mathbf{P}(\mathit{Dolor}, \mathit{Hueco}, \mathit{Caries}) =$$

$$\mathbf{P}(\mathit{Dolor}, \mathit{Hueco}|\mathit{Caries})\mathbf{P}(\mathit{Caries}) =$$

$$= \mathbf{P}(\mathit{Dolor}|\mathit{Caries})\mathbf{P}(\mathit{Hueco}|\mathit{Caries})\mathbf{P}(\mathit{Caries})$$

  - En lugar de tener una tabla con 7 números independientes sólo necesitamos 5 números independientes (en tres tablas)

Independencia condicional

- Si *Caries* tuviera n efectos independientes entre sí (dado *Caries*), el tamaño de la representación de la DCC crece $O(n)$ en lugar de $O(2^n)$
- En general, con una *Causa* con n efectos E_i independientes entre sí dado *Causa*, se tiene

$$\mathbf{P}(Causa, E_1, \ldots, E_n) = \mathbf{P}(Causa) \prod_i \mathbf{P}(E_i | Causa)$$

- No siempre se dan estas condiciones de independencia tan fuertes, aunque a veces compensa asumirlas
- En general, las *relaciones causa-efecto* y las de *independiencia condicional* pueden darse a varios niveles entre las v. a. de un dominio. El conocer estas relaciones es fundamental para representar de manera eficiente la DCC.

El ejemplo de la alarma (Pearl, 1990):

- 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: por tener música muy alta en su casa, por ejemplo
  - Incluso podrían llamar aunque no hubiera sonado: por confundirla con un teléfono, por ejemplo

Modelando el ejemplo de la alarma

- La indertidumbre que subyace hace aconsejable un modelo probabilístico
- Variables aleatorias (todas booleanas):
  - *Robo*: se ha producido un robo
  - *Terremoto*: ha ocurrido un terremoto
  - *Alarma*: la alarma ha sonado
  - *Juan*: Juan llama a la policía
  - *Maria*: María llama a la policía
- Relaciones de independencia:
  - *Robo* y *Terremoto* son independientes
  - Dado *Alarma*, *Juan* es condicionalmente independiente del resto de variables. Lo mismo para *Maria*.
    - Es decir, sabiendo si la alarma ha sonado o no, nuestro grado de creencia de en el hecho de que Juan (o María) llame a la policía, no se verá actualizado si además sabemos el valor de alguna de las otras variables.

Factorizando la DCC en el ejemplo de la alarma

- Aplicando la regla del producto repetidas veces en un orden determinado, y las relaciones de independencia condicional, tenemos:

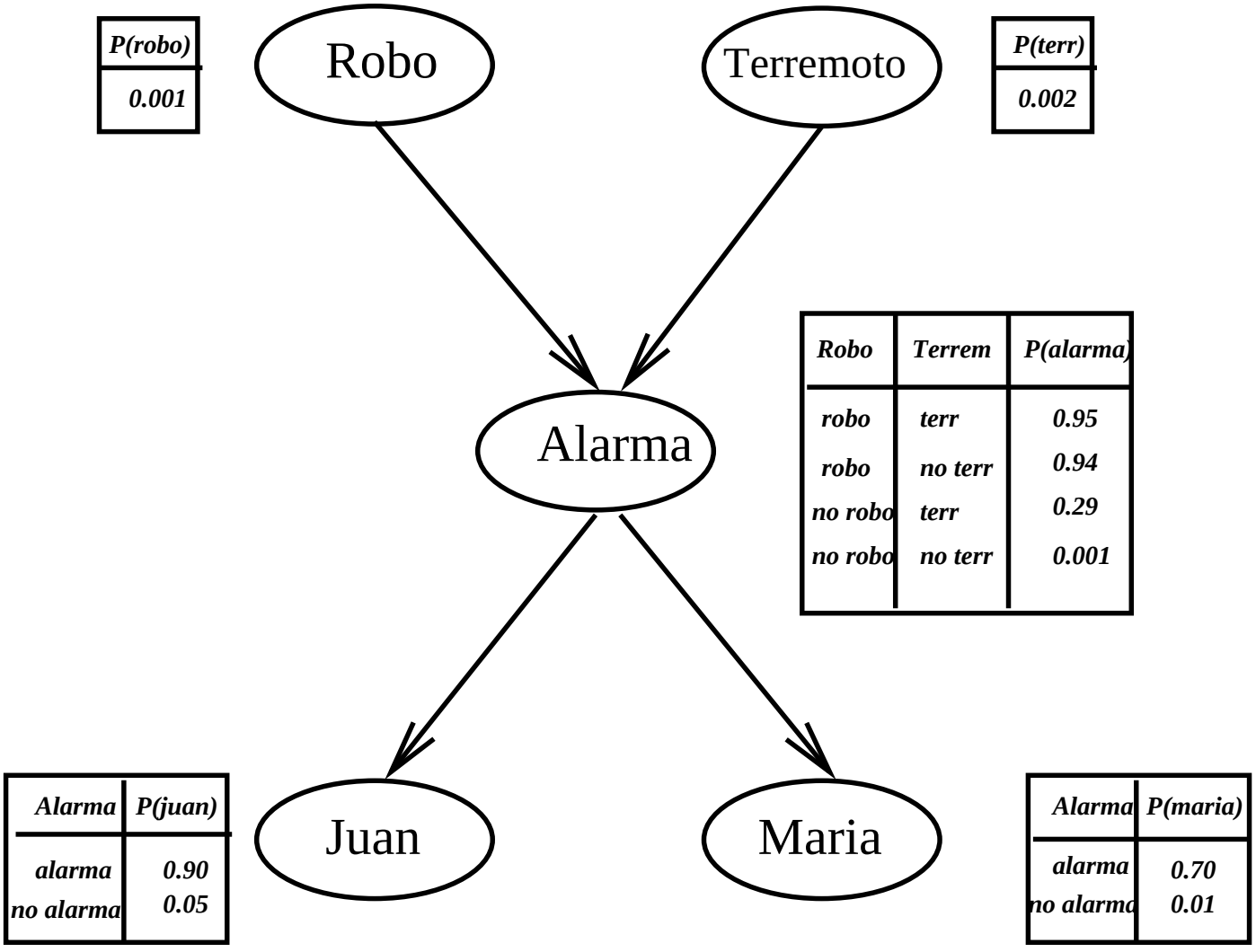
$$\mathbf{P}(Juan, Maria, Alarma, Robo, Terremoto) =$$
$$= \mathbf{P}(Juan|Maria, Alarma, Robo, Terremoto)\cdot\mathbf{P}(Maria|Alarma, Robo, Terremoto)\cdot$$
$$\cdot\mathbf{P}(Alarma|Robo, Terremoto)\cdot\mathbf{P}(Robo|Terremoto)\cdot\mathbf{P}(Terremoto) =$$
$$= \mathbf{P}(Juan|Alarma)\cdot\mathbf{P}(Maria|Alarma)\cdot\mathbf{P}(Alarma|Robo, Terremoto)\cdot$$
$$\cdot\mathbf{P}(Robo)\cdot\mathbf{P}(Terremoto)$$

- Observaciones:
  - En lugar de almacenar la DCC completa, basta con almacenar cinco tablas más pequeñas
  - Las probabilidades que se necesitan saber son condicionales, de tipo causa-efecto, que son las que normalmente se conocen a priori

Redes Bayesianas

- En general, las relaciones de independencia condicional permiten simplificar drásticamente las DCCs, haciendo que se puedan usar en la práctica
- Las *redes bayesianas* (o *redes de creencia*) constituyen una manera práctica y compacta de representar el conocimiento incierto, basada en esta idea

Red bayesiana para el ejemplo de la alarma (J. Pearl)



Redes bayesianas

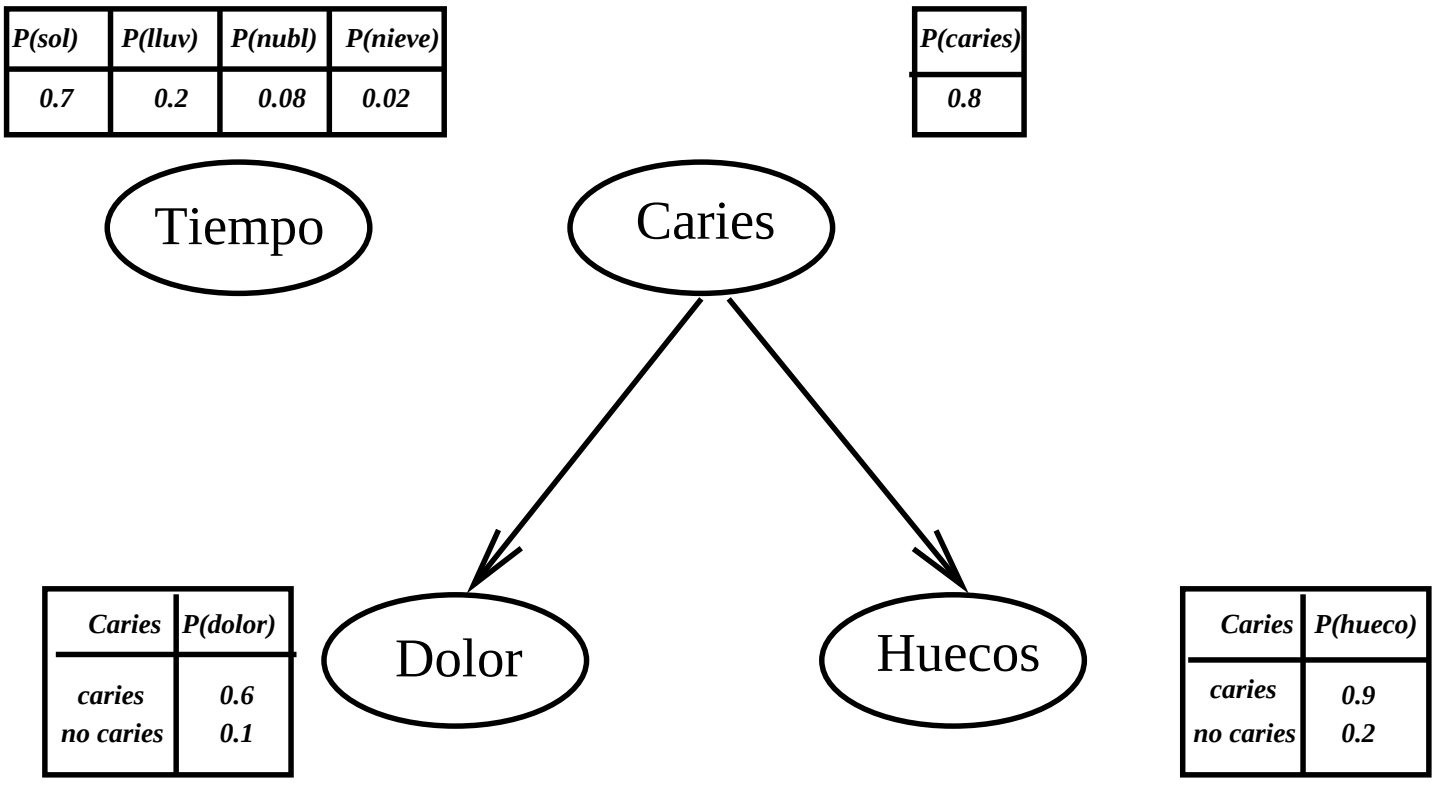
Una red bayesiana es un grafo dirigido *acíclico* que consta de:

- *Un conjunto de nodos*, uno por cada variable aleatoria del "mundo" que se representa
- *Un conjunto de arcos dirigidos* que conectan los nodos
  - Si hay un arco de X a Y decimos que X es un *padre* de Y ($padres(X)$ denota el conjunto de v.a. que son padres de X).
  - A partir del concepto de variable padre, se definen también los *antecesores* y *descendientes* de una variable.
- Cada nodo X_i contiene la *distribución de probabilidad condicional* $\mathbf{P}(X_i|padres(X_i))$
  - Si X_i es booleana, se suele omitir la probabilidad de la negación.

Redes bayesianas: intuición

- Intuitivamente, si se conoce el valor que toman las variables padre de una variable aleatoria, entonces el grado de creencia en que la variable tome un valor determinado, no se vería actualizado si además conociéramos el valor que toma alguna otra variable no descendiente.
- Es tarea del experto en el dominio (o de aprendizaje automático) el decidir las relaciones de independencia condicional (es decir, la *topología* de la red)

Ejemplo de la caries como red bayesiana (Russell y Norvig)



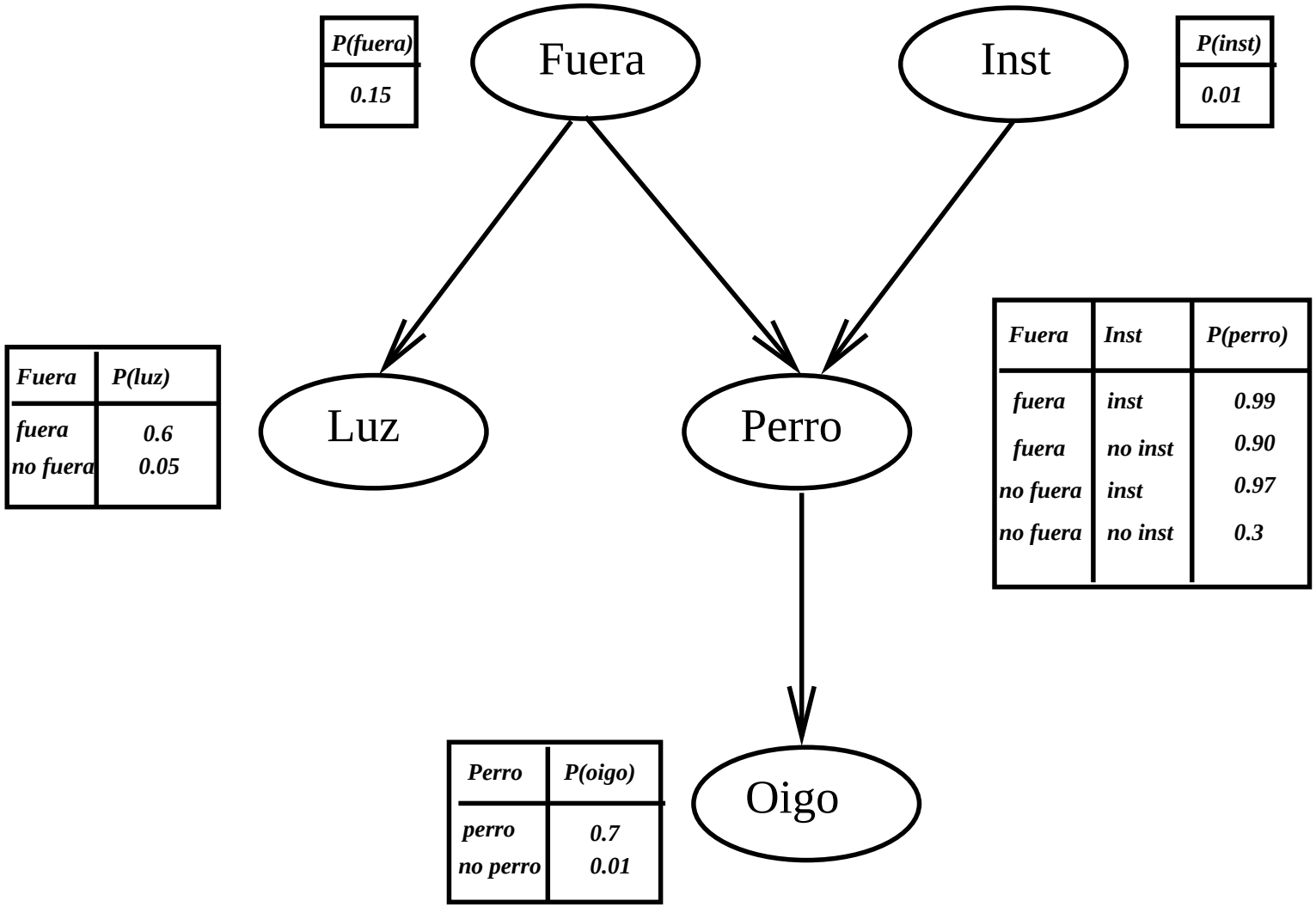
Observaciones sobre el ejemplo

- La topología de la red anterior nos expresa que:
  - *Caries* es una *causa directa* de *Dolor* y *Huecos*
  - *Dolor* y *Huecos* son condicionalmente independientes dada *Caries*
  - *Tiempo* es independiente de las restantes variables
- No es necesario dar la probabilidad de las negaciones de *caries*, *dolor*, ...

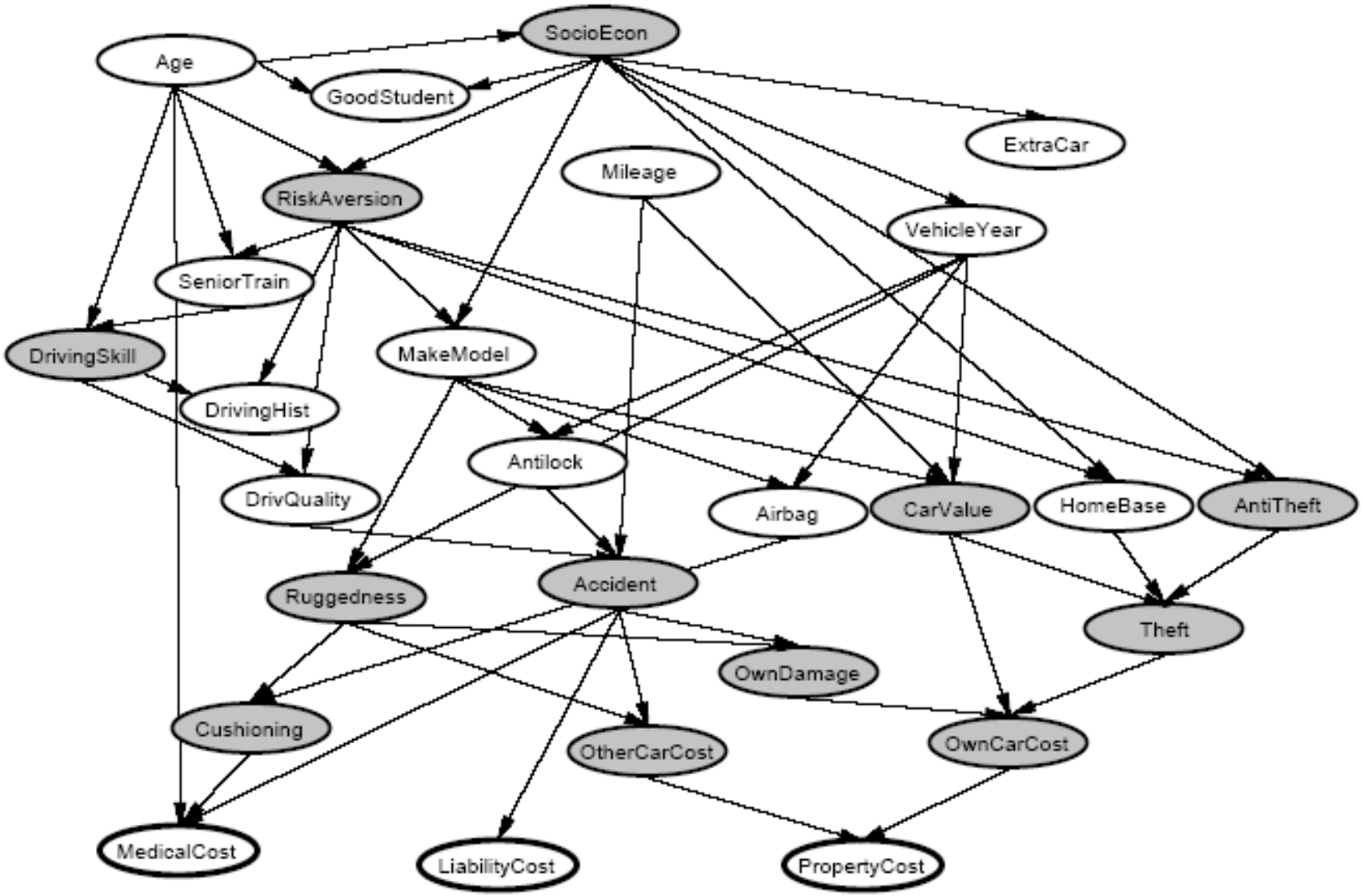
Ejemplo de la familia fuera de casa (Charniak, 1991):

- 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 ladrar y pensar que es mi perro aunque no fuera así
- Variables aleatorias (booleanas) en este problema:
  - *Fuera* (nadie en casa), *Luz* (luz en la entrada), *Perro* (perro fuera), *Inst* (problemas intestinales en el perro) y *Oigo* (oigo al perro ladrar)

Red bayesiana para el ejemplo de la familia fuera de casa



Ejemplo del seguro de automóvil (Binder et al.)



Las redes bayesianas representan DCCs

- Consideremos una red bayesiana con n variables aleatorias
  - Y un orden entre esas variables: $X_1, \ldots, X_n$
- En lo que sigue, supondremos que:
  - $padres(X_i) \subseteq \{X_{i-1}, \ldots, X_1\}$ (para esto, basta con que el orden escogido sea consistente con el orden parcial que induce el grafo)
  - $\mathbf{P}(X_i|X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i|padres(X_i))$ (es decir, cada variable es condicionalmente independiente de todos los que no son sus descendientes, dados sus padres en la red)
- Estas condiciones expresan formalmente nuestra intuición al representar nuestro "mundo" mediante la red bayesiana correspondiente
  - Por ejemplo, la red de la alarma expresa que creemos que $\mathbf{P}(Maria|Juan, Alarma, Terremoto, Robo) = \mathbf{P}(Maria|Alarma)$

Las redes bayesianas representan DCCs

- En las anteriores condiciones, y aplicando repetidamente la regla del producto:

$$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_n | X_{n-1} \ldots, X_1)\mathbf{P}(X_{n-1} \ldots, X_1) = \ldots$$

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

- Una red bayesiana *representa una DCC* obtenida mediante la expresión $\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | padres(X_i))$
  - En el ejemplo de la alarma, la probabilidad de que la alarma suene, Juan y María llamen a la policía, pero no haya ocurrido nada es (usamos iniciales, por simplificar):

$$P(j, m, a, \neg r, \neg t) = P(j|a)P(m|a)P(a|\neg r, \neg t)P(\neg r)P(\neg t) =$$

$$= 0{,}9 \times 0{,}7 \times 0{,}001 \times 0{,}999 \times 0{,}998 = 0{,}00062$$

Representaciones compactas

- Dominios *localmente estructurados*:
  - Las relaciones de independencia que existen entre las variables de un dominio hacen que las redes bayesianas sean una representación mucho más compacta y eficiente de una DCC que la tabla con todas las posibles combinaciones de valores
- Además, para un experto en un dominio de conocimiento suele ser más natural dar probabilidades condicionales que directamente las probabilidades de la DCC

Representaciones compactas

- Con n variables, si cada variable está directamente influenciada por k variables a lo sumo, entonces una red bayesiana necesitaría $n \cdot 2^k$ números, frente a los 2^n números de la DCC
  - Por ejemplo, Para $n = 30$ y $k = 5$, esto supone 960 números frente a 2^{30} (billones)
- Hay veces que una variable influye directamente sobre otra, pero esta dependencia es muy tenue
  - En ese caso, puede compensar no considerar esa dependencia, perdiendo algo de precisión en la representación, pero ganando manejabilidad

Redes bayesianas: problemas a tratar

En lo que queda de tema, veremos las siguientes cuestiones relacionadas con redes bayesianas:

- Creación de la red de manera que modelice bien la realidad representada y que sea compacta
- Deducción de independencias condicionales entre grupos de variables
  - *Deducir* independencias.
  - Criterio gráfico de *d-separación*
- Inferencia probabilística (cáculo de probabilidades del tipo $\mathbf{P}(X|\mathbf{e})$
  - Inferencia exacta: algoritmos de enumeración y de eliminación de variables
  - Inferencia aproximada: algoritmos de muestreo con rechazo y de ponderación por verosimilitud

Algoritmo de construcción de una red bayesiana

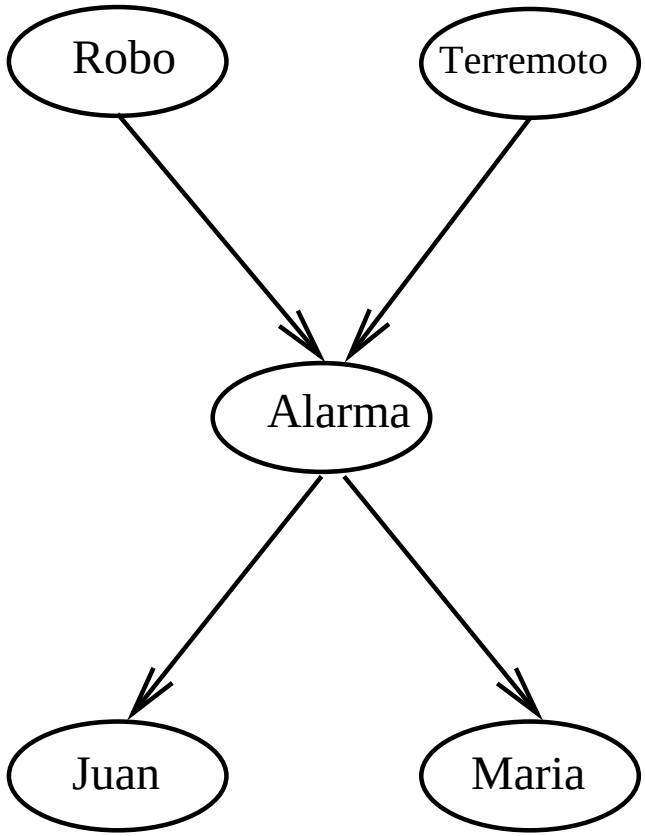
- Supongamos dado un conjunto de variables aleatorias **VARIABLES** que representan un dominio de conocimiento (con incertidumbre)

Algoritmo de construcción de redes bayesianas

FUNCION CONSTRUYE_RED(VARIABLES)
1. Sea $(X_1, \ldots, X_n)$ **una ordenación de VARIABLES**
2. Sea RED una red bayesiana ``vacía''
3. PARA $i = 1, \ldots, n$ **HACER**
  3.1 Añadir un nodo etiquetado con X_i **a RED**
  3.2 Sea padres(X_i**) un subconjunto minimal de** $\{X_{i-1}, \ldots, X_1\}$
    tal que existe una independencia condicional entre X_i
    y cada elemento de $\{X_{i-1}, \ldots, X_1\}$ **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

Ejemplo de construcción de red bayesiana (alarma)

- Partiendo del orden *Robo*, *Terremoto*, *Alarma*, *Juan*, *Maria*, y aplicando el algoritmo anterior obtenemos la red del ejemplo:



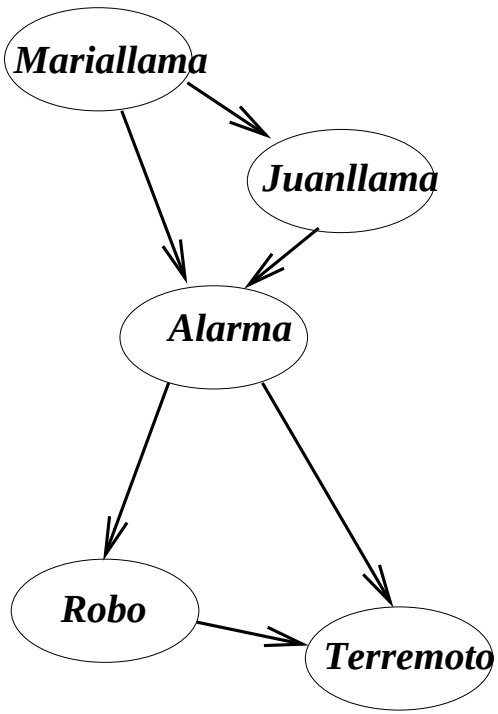
.

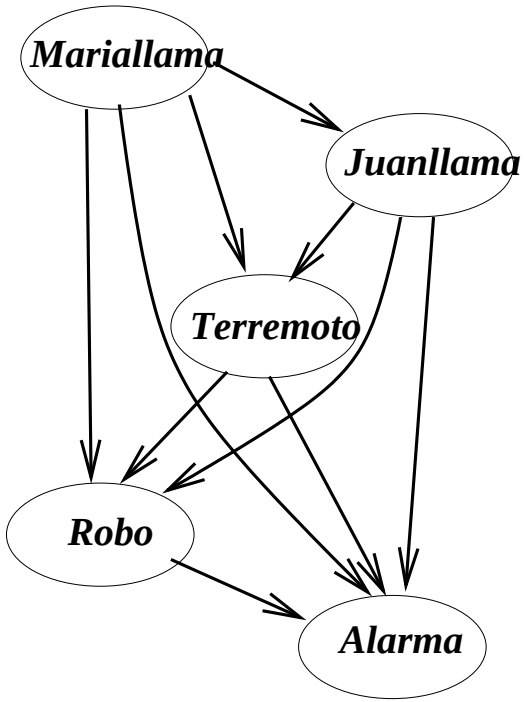
Construcción de redes bayesianas

- Problema: elección del orden entre variables
  - 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*)
  - Esto hará que las tablas reflejen probabilidades "causales" más que "diagnósticos", lo cual suele ser preferible por los expertos

Construcción de redes bayesianas

- Un orden malo puede llevar a representaciones poco eficientes
- Ejemplo: red izquierda (*Maria*, *Juan*, *Alarma*, *Robo* y *Terremoto*) y red derecha (*Maria*, *Juan*, *Terremoto*, *Robo* y *Alarma*)



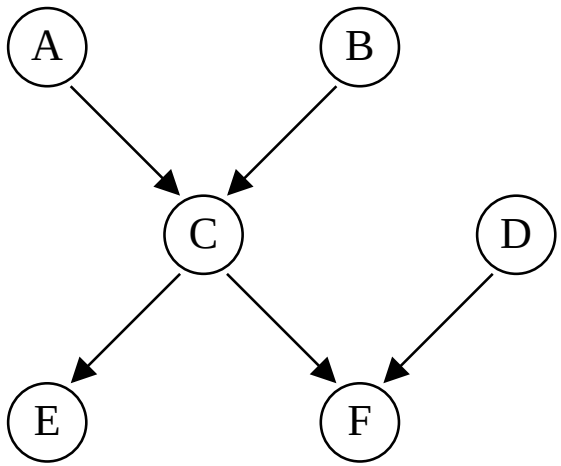


Independencia condicional en redes bayesianas

- Una vez dibujada la red (asumiendo una serie de independencias condicionales), de la propia estructura de la red se pueden *deducir* otras relaciones de independencia condicional
  - El concepto de *d-separación* nos proporciona un criterio gráfico para deducir estas independencias condicionales
- En una red bayesiana, un conjunto de variables aleatorias $\mathcal{Z}$ *d-separa* dos variables aleatorias $\mathbf{X}$ e $\mathbf{Y}$ si y sólo si todo camino en la red que una $\mathbf{X}$ con $\mathbf{Y}$, cumple alguna de estas condiciones:
  - Pasa por un elemento $\mathbf{Z} \in \mathcal{Z}$ con arcos en el mismo sentido $(\ldots \longleftarrow \mathbf{Z} \longleftarrow \ldots, \ldots \longrightarrow \mathbf{Z} \longrightarrow \ldots)$ o divergentes $(\ldots \longleftarrow \mathbf{Z} \longrightarrow \ldots)$
  - Pasa por un elemento $\mathbf{Z} \notin \mathcal{Z}$ con arcos convergentes $(\ldots \longrightarrow \mathbf{Z} \longleftarrow \ldots)$ y ningún descendiente de $\mathbf{Z}$ está en $\mathcal{Z}$
- Teorema: En una red bayesiana, dos variables aleatorias son condicionalmente independientes dado cualquier conjunto de variables aleatorias que las d-separe

Independencia condicional en redes bayesianas

- Ejemplos



  - A es (condicionalmente) independiente de B dado $\emptyset$
  - E es condicionalmente independiente de F dado $\{C\}$
  - $\{F\}$ no d-separa a A y D
  - A es condicionalmente independiente de D dado $\{C\}$
  - $\{F\}$ no d-separa a E y D.
  - D es condicionalmente independiente de E dado $\{C\}$

Inferencia probabilística en una red bayesiana

- El problema de la inferencia en una red bayesiana:
  - 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, podríamos querer saber qué probabilidad hay de que realmente se haya producido un robo, sabiendo que tanto Juan como María han llamado a la policía
  - Es decir, calcular $\mathbf{P}(Robo|juan, maria)$

Inferencia probabilística en una red bayesiana

- Notación:
  - X denotará la variable de consulta (sin pérdida de generalidad supondremos sólo una variable)
  - $\mathbf{E}$ denota un conjunto de *variables de evidencia* $E_1, E_2, \ldots, E_n$ y $\mathbf{e}$ una observación concreta para esas variables
  - $\mathbf{Y}$ denota al conjunto de las restantes variables de la red e $\mathbf{y}$ representa un conjunto cualquiera de valores para esas variables

Inferencia por enumeración

- Recordar la fórmula para la inferencia probabilística a partir de una DCC:

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

- Esta fórmula será la base para la inferencia probabilística:
  - Puesto que una red bayesiana es una representación de una DCC, nos permite calcular cualquier probabilidad a posteriori a partir de la información de la red bayesiana
  - Esencialmente, se trata de una suma de productos de los elementos de las tablas de las distribuciones condicionales

Un ejemplo de inferencia probabilística

- Ejemplo de la alarma (usamos iniciales por simplificar):

$$\mathbf{P}(R|j,m) = \langle P(r|j,m); P(\neg r|j,m)\rangle =$$

$$= \alpha\langle \sum_t \sum_a P(r,t,a,j,m); \sum_t \sum_a P(\neg r,t,a,j,m)\rangle =$$

$$= \alpha\langle \sum_t \sum_a P(r)P(t)P(a|r,t)P(j|a)P(m|a);$$

$$\sum_t \sum_a P(\neg r)P(t)P(a|\neg r,t)P(j|a)P(m|a)\rangle$$

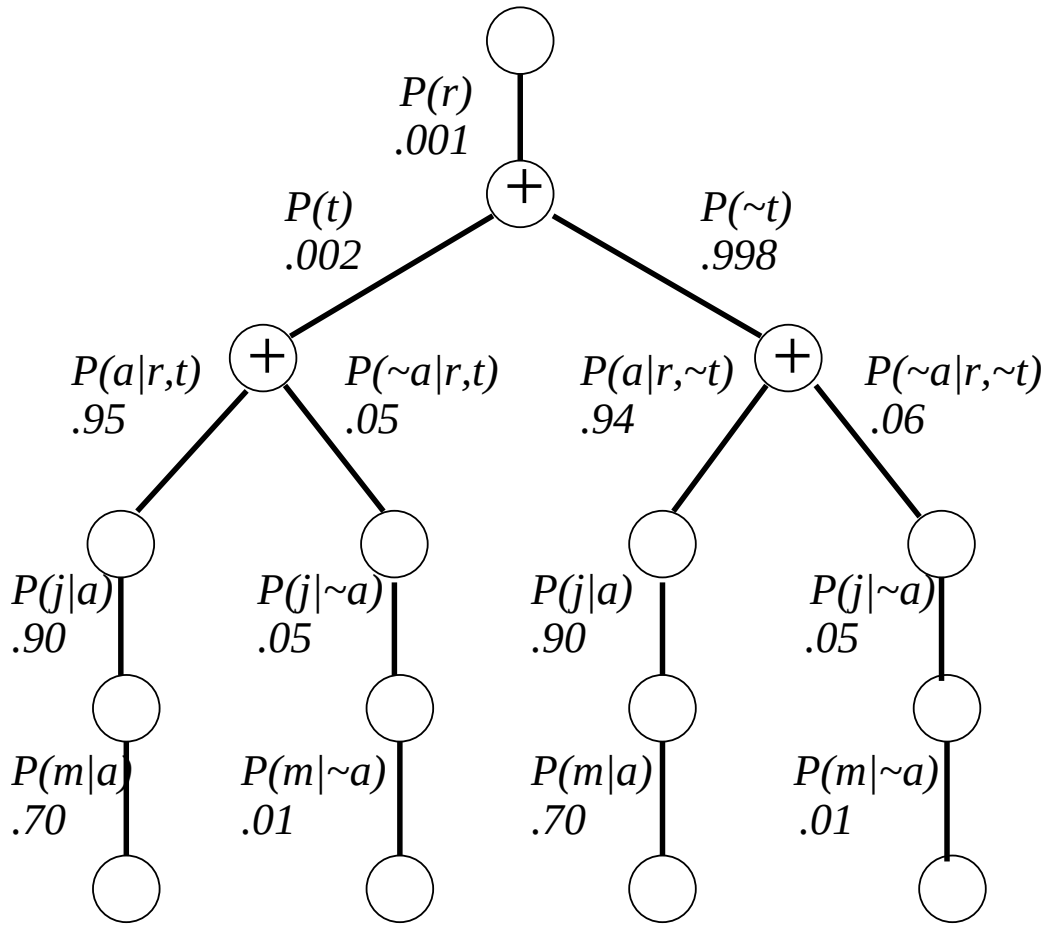
Un ejemplo de inferencia probabilística

- En este ejemplo hay que hacer 2×4 sumas, cada una de ellas con un producto de cinco números tomados de la red bayesiana
  - En el peor de los casos, con n variables booleanas, este cálculo toma $O(n2^n)$
- Una primera mejora consiste en sacar factor común de aquellas probabilidades que sólo involucran variables que no aparecen en el sumatorio:

$$\mathbf{P}(R|j,m) = \alpha\langle P(r)\sum_t P(t)\sum_a P(a|r,t)P(j|a)P(m|a);$$
$$P(\neg r)\sum_t P(t)\sum_a P(a|\neg r,t)P(j|a)P(m|a)\rangle =$$
$$= \alpha\langle 0{,}00059224; 0{,}0014919\rangle = \langle 0{,}284; 0{,}716\rangle$$

Inferencia por enumeración

- Las operaciones realizadas en la fórmula anterior se pueden simbolizar con el siguiente árbol:



Algoritmo de inferencia por enumeración

- Entrada: una v.a. X de consulta, un conjunto de valores observados $\mathbf{e}$ para la variables de evidencia y una red bayesiana
- Salida: $\mathbf{P}(X|\mathbf{e})$

Algoritmo de inferencia por enumeración

```
FUNCION INFERENCIA_ENUMERACION(X,e,RED)
1. Sea Q(X) una distribución de probabilidad sobre X,
   inicialmente vacía
2. PARA cada valor x_i de X HACER
   2.1 Extender e con el valor x_i para X
   2.2 Hacer Q(x_i) el resultado de
       ENUM_AUX(VARIABLES(RED),e,RED)
3. Devolver NORMALIZA(Q(X))
```

Algoritmo de inferencia por enumeración

Algoritmo de inferencia por enumeración

FUNCION ENUM_AUX(VARS,e,RED)
1. Si VARS es vacío devolver 1
2. Si no,
2.1 Hacer Y igual a PRIMERO(VARS)
2.2 Si Y tiene un valor y en e,
2.2.1 devolver P(y\|padres(Y,e))·ENUM_AUX(RESTO(VARS),e)
2.2.2 Si no, devolver
SUMATORIO(y,P(y\|padres(Y,e))·ENUM_AUX(RESTO(VARS),e_y))
(donde: padres(Y,e) es el conjunto de valores que toman en e los padres de Y en la RED, y e_y extiende e con el valor y para Y)

Algoritmo de inferencia por enumeración

- Observación:
  - Para que el algoritmo funcione, **`VARIABLES(RED)`** debe devolver las variables en un orden consistente con el orden implícito en el grafo de la red (de arriba hacia abajo)
- Recorrido en profundidad:
  - El algoritmo genera el árbol de operaciones anterior de arriba hacia abajo, en profundidad
  - Por tanto, tiene un coste lineal en espacio
- Puede realizar cálculos repetidos
  - En el ejemplo, $P(j|a)P(m|a)$ y $P(j|\neg a)P(m|\neg a)$ se calculan dos veces
  - Cuando hay muchas variables, estos cálculos redundantes son inaceptables en la práctica

Evitando cálculos redundantes

- Idea para evitar el cálculo redundante:
  - Realizar las operaciones correspondientes a cada sumatorio *una sóla vez*, para todas los posibles valores de las variables que intervienen en ese sumatorio
  - En lugar de multiplicar *números*, multiplicaremos *tablas de probabilidades*
  - Denominaremos *factores* a estas tablas

Evitando cálculos redundantes

- Por ejemplo, la operación

$$\mathbf{P}(R|j,m) = \alpha \mathbf{P}(R) \sum_t P(T) \sum_a \mathbf{P}(A|R,T) P(j|A) P(m|A)$$

  puede verse como una serie de operaciones entre cinco tablas o *factores*
- Estas operaciones entre tablas son de dos tipos: multiplicación (componente a componente) o agrupación (sumando para los dstintos valores de una variable)
- Es el denominado algoritmo de *eliminación de variables*
- Veremos con más detalle cómo actuaría este algoritmo para calcular $\mathbf{P}(R|j,m)$

El algoritmo de eliminación de variables: un ejemplo

- En primer lugar, tomamos los factores iniciales que intervienen en el cálculo
  - Estos factores salen directamente de las tablas de probabilidad de la red,
  - Cada variable aporta su propio factor, el que corresponde a su tabla de probabilidad en la red.
  - En esas tablas, los valores de las variables de evidencia están fijados

El algoritmo de eliminación de variables: un ejemplo

- El factor correspondiente a M se obtiene a partir de la distribución condicional $\mathbf{P}(M|A)$.
  - Como M es una variable de evidencia y su valor está fijado a *true*, el factor correspondiente, notado $\mathbf{f_M}(A)$, es $\mathbf{P}(m|A)$ (tabla con componentes $P(m|a)$ y $P(m|\neg a)$):

A	$\mathbf{f_M}(A) = \mathbf{P}(m\|A)$
a	0,70
$\neg a$	0,01

- El factor correspondiente a J se obtiene a partir de la distribución condicional $\mathbf{P}(J|A)$
  - Como J es una variable de evidencia y su valor está fijado a *true*, el factor correspondiente (notado $\mathbf{f_J}(A)$), es $\mathbf{P}(j|A)$:

A	$\mathbf{f_J}(A) = \mathbf{P}(j\|A)$
a	0,90
$\neg a$	0,05

El algoritmo de eliminación de variables: un ejemplo

- El factor correspondiente a la variable A, notado $\mathbf{f_A}(A, R, T)$ se obtiene a partir de $\mathbf{P}(A|R, T)$
  - Ninguna de esas variables es de evidencia, y por tanto no están fijados sus valores
  - Es una tabla con $2 \times 2 \times 2$ entradas, una por cada combinación de valores de A, R y T
  - En este caso, esta tabla está directamente en la propia red

A	R	T	$\mathbf{f_A}(A, R, T) = \mathbf{P}(A\|R, T)$
a	r	t	0,95
a	r	$\neg t$	0,94
a	$\neg r$	t	0,29
a	$\neg r$	$\neg t$	0,001
$\neg a$	r	t	0,05
$\neg a$	r	$\neg t$	0,06
$\neg a$	$\neg r$	t	0,71
$\neg a$	$\neg r$	$\neg t$	0,999

El algoritmo de eliminación de variables: un ejemplo

- El factor correspondiente a la variable T, que notaremos $\mathbf{f_T}(T)$, es la tabla $\mathbf{P}(T)$

T	$\mathbf{f_T}(T) = \mathbf{P}(T)$
t	0,002
$\neg t$	0,998

- El factor correspondiente a R, que notaremos $\mathbf{f_R}(R)$, es la tabla $\mathbf{P}(R)$

R	$\mathbf{f_R}(R) = \mathbf{P}(R)$
r	0,001
$\neg r$	0,999

El algoritmo de eliminación de variables: un ejemplo

- Una vez tenemos todos los factores iniciales vamos "eliminando" las variables que no son ni de consulta ni de evidencia.
  - Intuitivamente, el proceso de "eliminación" de una variable se corresponde con realizar el sumatorio de la correspondiente fórmula, pero a nivel de tablas.
  - Cuando eliminemos todas las variables (que no sean de evidencia ni de consulta), sólo nos quedarán factores dependientes de la variable de consulta, que habrán de ser multiplicados y finalmente normalizados.
- La eliminación de las variables la podemos realizar en cualquier orden sin que afecte al resultado final
  - Aunque el orden en el que se vayán eliminando las variables sí influirá en la eficiencia del algoritmo.

El algoritmo de eliminación de variables: un ejemplo

- Por ejemplo, eliminamos en primer lugar la variable A
  - Se correspondería con realizar $\sum_a \mathbf{P}(A|R,T)P(j|A)P(m|A)$, o lo que es lo mismo, hacer $\sum_a \mathbf{f_A}(A,R,T)\mathbf{f_J}(A)\mathbf{f_M}(A)$
  - La multiplicación de $\mathbf{f_M}$, $\mathbf{f_J}$ y $\mathbf{f_A}$, notada $\mathbf{f}_{\times\mathbf{A}}(A,R,T)$ se obtiene multiplicando las entradas correspondientes a los mismos valores de A, R y T
  - Es decir, para cada valor v_1 de A, v_2 de R y v_3 de T se tiene $\mathbf{f}_{\times\mathbf{A}}(v_1,v_2,v_3) = \mathbf{f_M}(v_1)\mathbf{f_J}(v_1)\mathbf{f_A}(v_1,v_2,v_3)$. Por ejemplo: $\mathbf{f}_{\times\mathbf{A}}(\mathit{true},\mathit{false},\mathit{true}) = \mathbf{f_M}(\mathit{true})\mathbf{f_J}(\mathit{true})\mathbf{f_A}(\mathit{true},\mathit{false},\mathit{true}) = 0{,}70 \times 0{,}90 \times 0{,}29 = 0{,}1827$

El algoritmo de eliminación de variables: un ejemplo

- Ahora hay que *agrupar* el valor de A en $\mathbf{f}_{\times \mathbf{A}}$ (realizar el sumatorio $\sum_a$)
  - Así, obtenemos una tabla $\mathbf{f}_{\overline{\mathbf{A}}}(R, T)$ haciendo $\mathbf{f}_{\overline{\mathbf{A}}}(v_1, v_2) = \sum_a \mathbf{f}_{\times \mathbf{A}}(a, v_1, v_2)$ para cada valor v_1 de R y v_2 de T, y variando a en los posibles valores de A
  - Llamaremos a esta operación *agrupamiento*
  - Hemos *eliminado* la variable A
  - Una vez realizada la agrupación, guardamos $\mathbf{f}_{\overline{\mathbf{A}}}$ y nos olvidamos de $\mathbf{f}_{\mathbf{M}}, \mathbf{f}_{\mathbf{J}}$ y $\mathbf{f}_{\mathbf{A}}$

R	T	a	$\neg a$	$\mathbf{f}_{\overline{\mathbf{A}}}(R, T)$
r	t	$0{,}70 \times 0{,}90 \times 0{,}95 = 0{,}5985$	$0{,}01 \times 0{,}05 \times 0{,}05 = 0{,}00003$	0,59853
r	$\neg t$	$0{,}70 \times 0{,}90 \times 0{,}94 = 0{,}5922$	$0{,}01 \times 0{,}05 \times 0{,}06 = 0{,}00003$	0,59223
$\neg r$	t	$0{,}70 \times 0{,}90 \times 0{,}29 = 0{,}1827$	$0{,}01 \times 0{,}05 \times 0{,}71 = 0{,}00036$	0,18306
$\neg r$	$\neg t$	$0{,}70 \times 0{,}90 \times 0{,}001 = 0{,}00063$	$0{,}01 \times 0{,}05 \times 0{,}999 = 0{,}0005$	0,00113

El algoritmo de eliminación de variables: un ejemplo

- Eliminación de T:
  - Notemos por $\mathbf{f}_{\overline{\mathbf{T}}}(R)$ al resultado de multiplicar $\mathbf{f}_{\overline{\mathbf{A}}}(R, T)$ por $\mathbf{f}_{\mathbf{T}}(T)$ y agrupar por T

R	t	$\neg t$	$\mathbf{f}_{\overline{\mathbf{T}}}(R)$
r	$0{,}59853 \times 0{,}002 = 0{,}001197$	$0{,}59223 \times 0{,}998 = 0{,}591046$	$0{,}59224$
$\neg r$	$0{,}18306 \times 0{,}002 = 0{,}000366$	$0{,}00113 \times 0{,}998 = 0{,}001128$	$0{,}00149$

  - Podemos olvidarnos de $\mathbf{f}_{\overline{\mathbf{A}}}$ y $\mathbf{f}_{\mathbf{T}}$

El algoritmo de eliminación de variables: un ejemplo

- Queda la variable R:
  - Al ser la variable de consulta, no se agrupa, sólo multiplicamos los factores en los que aparece
  - Multiplicamos $\mathbf{f}_{\overline{\mathbf{T}}}(R)$ y $\mathbf{f}_{\mathbf{R}}(R))$ para obtener $\mathbf{f}_{\times \mathbf{R}}(R)$

R	$\mathbf{f}_{\overline{\mathbf{T}}}(R) \times \mathbf{f}_{\mathbf{R}}(R)$
r	$0{,}59224 \times 0{,}001 = 0{,}00059$
$\neg r$	$0{,}00149 \times 0{,}999 = 0{,}00149$

  - Finalmente, normalizamos la tabla (para que sus componentes sumen 1): $\mathbf{P}(R|j,m) = \langle 0{,}28417; 0{,}71583 \rangle$
- La tabla finalmente devuelta es la distribución $\mathbf{P}(R|j,m)$. Es decir, $P(r|j,m) = 0{,}28417$ y $P(\neg r|j,m) = 0{,}71583$

El algoritmo de eliminación de variables: observaciones

- Inicialmente, tenemos un conjunto de factores (uno por cada variable de la red) y vamos reduciendo el conjunto de factores:
- Por cada variable que no sea de consulta ni de evidencia, sustituimos los factores en los que aparece, por un único factor, resultante de "eliminar" la variable.
- Eliminar variable: multiplicar los factores en los que aparece y agrupar el factor producto (obteniendo un factor en el que ya no aparece la variable).
- Cuando se trata la variable de consulta, sólo se multiplican sus factores (no hay que agrupar).
- Al final del proceso, tendremos un único factor, que dependerá de la variable de consulta. Normalizando ese factor se obtiene el resultado de la inferencia probabilística.

El algoritmo de eliminación de variables: observaciones

- Multiplicación de tablas:
  - Si $\mathbf{f_1}(\mathbf{X}, \mathbf{Y})$ y $\mathbf{f_2}(\mathbf{Y}, \mathbf{Z})$ son dos tablas cuyas variables en común son las de $\mathbf{Y}$, se define su producto $\mathbf{f}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ como la tabla cuyas entradas son $f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_1(\mathbf{x}, \mathbf{y}) f_2(\mathbf{y}, \mathbf{z})$
  - Similar a una operación *join* en bases de datos, multiplicando los valores correspondientes
- Agrupamiento de una tabla:
  - Si $\mathbf{f}_{\times Y}(Y, \mathbf{Z})$ es un factor (que ha resultado de multiplicar los factores en los que aparece Y), su agrupamiento es una tabla $\mathbf{f}_{\overline{Y}}(\mathbf{Z})$ cuyas entradas son $\mathbf{f}_{\overline{Y}}(\mathbf{z}) = \sum_y \mathbf{f}_{\times Y}(y, \mathbf{Z})$
  - La operación de sumar por un valor es similar a la *agregación* de una columna en bases de datos

Un paso previo de optimización: variables irrelevantes

- En el algoritmo de eliminación de variables, se suele realizar un paso previo de eliminación de variables irrelevantes para la consulta
- Ejemplo:
  - Si la consulta a la red del ejemplo es $\mathbf{P}(J|r)$, hay que calcular $\alpha\mathbf{P}(r)\sum_t P(t)\sum_a P(a|r,t)\mathbf{P}(J|a)\sum_m P(m|a)$
  - Pero $\sum_m P(m|a) = 1$, así que la variable M es irrelevante para la consulta
  - Siempre podemos eliminar cualquier variable que sea una hoja de la red y que no sea de consulta ni de evidencia
- Y en general, se puede demostrar que toda variable que no sea antecesor (en la red) de alguna de las variables de consulta o de evidencia, es irrelevante para la consulta y por tanto *puede ser eliminada*

El algoritmo de eliminación de variables

- Entrada: una v.a. X de consulta, un conjunto de valores observados $\mathbf{e}$ para la variables de evidencia y una red bayesiana
- Salida: $\mathbf{P}(X|\mathbf{e})$

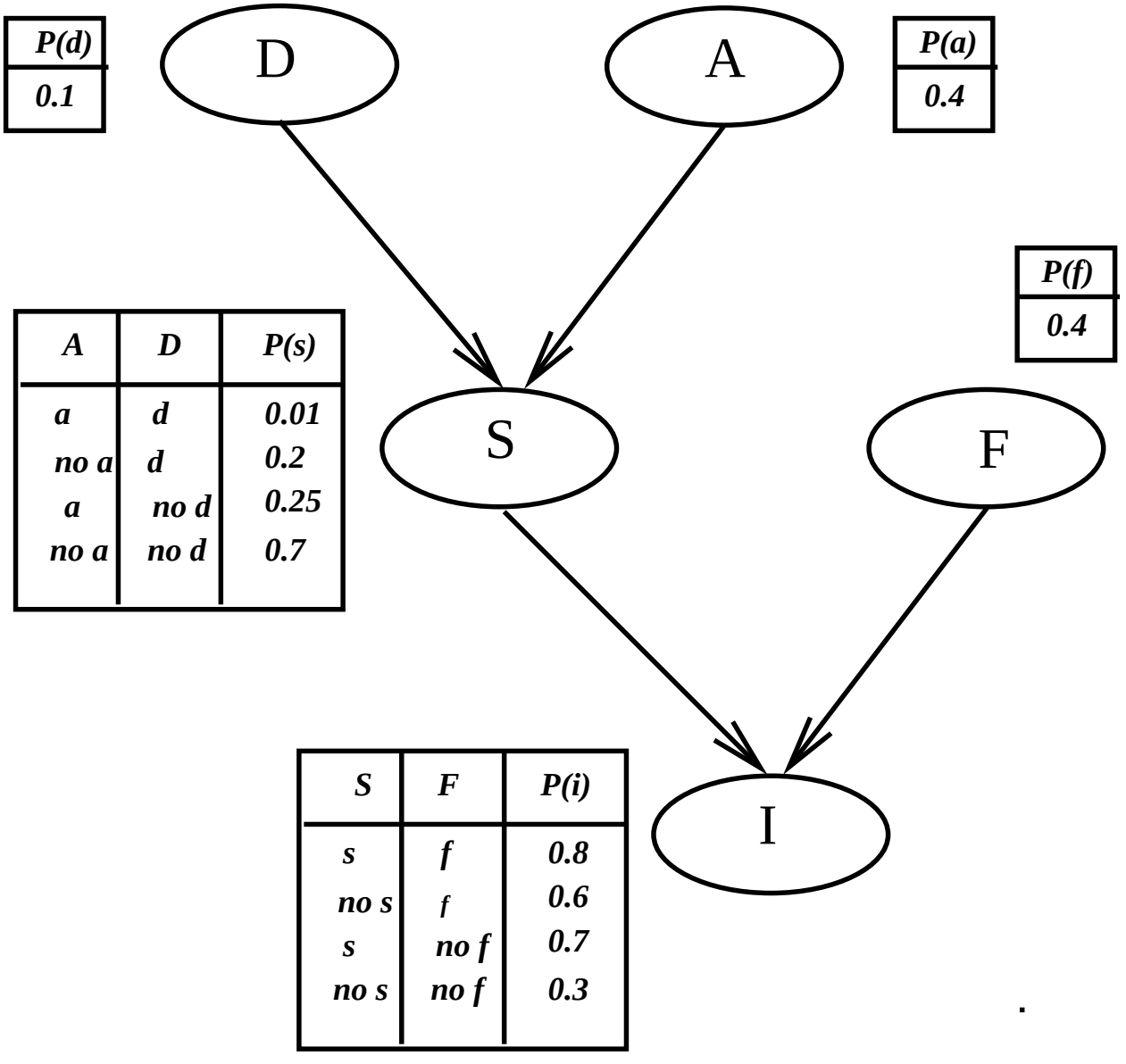
Algoritmo de eliminación de variables

```
FUNCION INFERENCIA_ELIMINACION_VARIABLES(X,e,RED)
1. Sea RED' el resultado de eliminar de RED las variables
   irrelevantes para la consulta realizada
2. Sea FACTORES igual a conjunto de los factores
   correspondientes a cada variable de RED'
4. Sea VARS_ORD el conjunto de las variables de RED' que
   no sean de evidencia, ordenado según un orden de
   eliminación
5. PARA cada VAR en VARS_ORD HACER
   5.1 Si VAR es de consulta, eliminar de FACTORES los
       factores en los que aparece VAR, e incluir el factor
       resultante de multiplicarlos
   5.3 Si VAR no es de consulta, eliminar de FACTORES los
       factores en los que aparece VAR, e incluir el factor
       resultante de  multiplicarlos y agruparlos por VAR
6. Devolver la NORMALIZACION del único factor
   que queda en FACTORES
```

Otro ejemplo (Béjar)

- 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

Ejemplo: red bayesiana



Ejemplo de inferencia probabilística

- Podemos usar la red bayesiana para calcular la probabilidad de ser fumador si se ha sufrido un infarto y no se hace deporte, $\mathbf{P}(F|i, \neg d)$
- Directamente:
  - Aplicamos la fórmula:
    $\mathbf{P}(F|i, \neg d) = \alpha \mathbf{P}(F, i, \neg d) = \alpha \sum_{S,A} \mathbf{P}(F, i, \neg d, A, S)$
  - Factorizamos según la red:
    $\mathbf{P}(F|i, \neg d) = \alpha \sum_{S,A} P(\neg d) P(A) P(S|\neg d, A) \mathbf{P}(F) \mathbf{P}(i|S, F)$
  - Sacamos factor común:
    $\mathbf{P}(F|i, \neg d) = \alpha P(\neg d) \mathbf{P}(F) \sum_{A} P(A) \sum_{S} P(S|\neg d, A) \mathbf{P}(i|S, F)$

Ejemplo de inferencia probabilística

- Calculemos:
  - Para $F = \mathit{true}$:

$$P(f|i,\neg d) = \alpha \cdot P(\neg d) \cdot P(f) \cdot$$

$$\cdot[P(a) \cdot (P(s|\neg d, a) \cdot P(i|s,f) + P(\neg s|\neg d, a) \cdot P(i|\neg s, f)) +$$

$$+P(\neg a) \cdot (P(s|\neg d, \neg a) \cdot P(i|s,f) + P(\neg s|\neg d, \neg a) \cdot P(i|\neg s, f))] =$$

$$= \alpha \cdot 0{,}9 \cdot 0{,}4 \cdot [0{,}4 \cdot (0{,}25 \cdot 0{,}8 + 0{,}75 \cdot 0{,}6) + 0{,}6 \cdot (0{,}7 \cdot 0{,}8 + 0{,}3 \cdot 0{,}6)] =$$

$$= \alpha \cdot 0{,}253$$

  - Análogamente, para $F = \mathit{false}$, $P(\neg f|i,\neg d) = \alpha \cdot 0,274$
  - Normalizando, $\mathbf{P}(F|i,\neg d) = \langle 0{,}48; 0{,}52 \rangle$

Aplicando eliminación de variables (factores inciales)

- Factor correspondiente a I:

S	F	$\mathbf{f_I}(S,F) = \mathbf{P}(i\|S,F)$
s	f	0,8
s	$\neg f$	0,7
$\neg s$	f	0,6
$\neg s$	$\neg f$	0,3

- Factor correspondiente a F:

F	$\mathbf{f_F}(F) = \mathbf{P}(F)$
f	0,4
$\neg f$	0,6

Aplicando eliminación de variables (factores iniciales)

- Factor correspondiente a S:

S	A	$\mathbf{f_S}(S, A) = \mathbf{P}(S \mid \neg d, A)$
s	a	0,25
s	$\neg a$	0,7
$\neg s$	a	0,75
$\neg s$	$\neg a$	0,3

- Factor correspondiente a A:

A	$\mathbf{f_A}(A) = \mathbf{P}(A)$
a	0,4
$\neg a$	0,6

- Factor correspondiente a D: $\mathbf{f_D}()$ (no depende de D, ya que su valor está fijado a $\neg d$, por tanto se trata de una tabla con una única entrada): 0,9
  - Como se verá, los factores de una sóla entrada en realidad son irrelevantes para el resultado final.

Aplicando eliminación de variables (orden de eliminación)

- Seguiremos el siguiente orden de variables, : S, A, F
- Puede que otro orden hiciera más eficiente el desarrollo del algoritmo
- Podríamos aplicar alguna heurística para elegir un buen orden de eliminación (ver más adelante)

Eliminación de S

- Primero multiplicamos $\mathbf{f_I}(S, F)$ y $\mathbf{f_S}(S, A)$ obteniendo $\mathbf{f}_{\times \mathbf{S}}(S, A, F)$ y en ese factor agrupamos por la variable S, obteniendo $\mathbf{f}_{\overline{\mathbf{S}}}(A, F)$:

A	F	s	$\neg s$	$\mathbf{f}_{\overline{\mathbf{S}}}(A, F)$
a	f	$0{,}25 \times 0{,}80$	$0{,}75 \times 0{,}60$	0,65
a	$\neg f$	$0{,}25 \times 0{,}70$	$0{,}75 \times 0{,}30$	0,40
$\neg a$	f	$0{,}70 \times 0{,}80$	$0{,}30 \times 0{,}60$	0,74
$\neg a$	$\neg f$	$0{,}70 \times 0{,}70$	$0{,}30 \times 0{,}30$	0,58

- Quitamos $\mathbf{f_I}(S, F)$ y $\mathbf{f_S}(S, A)$ de la lista de factores e incluimos $\mathbf{f}_{\overline{\mathbf{S}}}(A, F)$

Eliminación de A

- Primero multiplicamos $\mathbf{f_A}(A)$ y $\mathbf{f_{\overline{S}}}(A, F)$ obteniendo $\mathbf{f}_{\times \mathbf{A}}(A, F)$, y en ese factor agrupamos por la variable A, obteniendo $\mathbf{f_{\overline{A}}}(F)$

F	a	$\neg a$	$\mathbf{f_{\overline{A}}}(F)$
f	0,4,65	$0{,}6 \times 0{,}74$	0,704
$\neg f$	$0{,}4 \times 0{,}4$	$0{,}6 \times o{,}58$	0,508

- Quitamos $\mathbf{f_A}(A)$ y $\mathbf{f_{\overline{S}}}(A, F)$ de la lista de factores e incluimos $\mathbf{f_{\overline{A}}}(F)$

Aplicando eliminación de variables

- Ultimo paso: multiplicamos los factores restantes y normalizamos
  - Multiplicación

F	$\mathbf{f_D}() \times \mathbf{f_{\overline{A}}}(F) \times \mathbf{f_F}(F)$
f	$0{,}9 \times 0{,}704 \times 0{,}4 = 0{,}253$
$\neg f$	$0{,}9 \times 0{,}508 \times 0{,}6 = 0{,}274$

  - Normalizando obtenemos finalmente: $\mathbf{P}(F|i, \neg d) = \langle 0{,}48; 0{,}52 \rangle$ (obsérvese que en realidad el factor $\mathbf{f_D}$ era irrelevante, ya que no afecta al resultado de la normalización).
- Por tanto, la probabilidad de ser fumador, dado que se ha tenido un infarto y no se hace deporte, es del 48 %

Complejidad del algoritmo de eliminación de variables

- La complejidad del algoritmo (tanto en tiempo como en espacio) está dominada por el tamaño del mayor factor obtenido durante el proceso
- Y en eso influye el orden en el que se consideren las variables (*orden de eliminación*)
  - Podríamos usar un criterio heurístico para elegir el orden de eliminación
  - En general, estos criterios deciden *en cada paso* la siguiente variable a eliminar
  - Tratando de que los factores que vayan saliendo sean de un menor tamaño
- Si la red está *simplemente conectada* (*poliárbol*) se puede probar que la complejidad del algoritmo (en tiempo y espacio) es *lineal* en el tamaño de la red (el número de entradas en sus tablas)
  - Una red está simplemente conectada si hay a lo sumo un camino (no dirigido) entre cada dos nodos
  - Por ejemplo, la red del *Robo* está simplemente conectada

Complejidad de la inferencia exacta

- Pero en general, el algoritmo tiene complejidad exponencial (en tiempo y espacio) en el peor de los casos
  - Esto no es sorpendente: en particular la inferencia bayesiana incluye a la inferencia proposicional.
- Cuando la inferencia exacta se hace inviable, es esencial usar métodos aproximados de inferencia
- Métodos estocásticos, basados en muestreos que simulan las distribuciones de probabilidad de la red

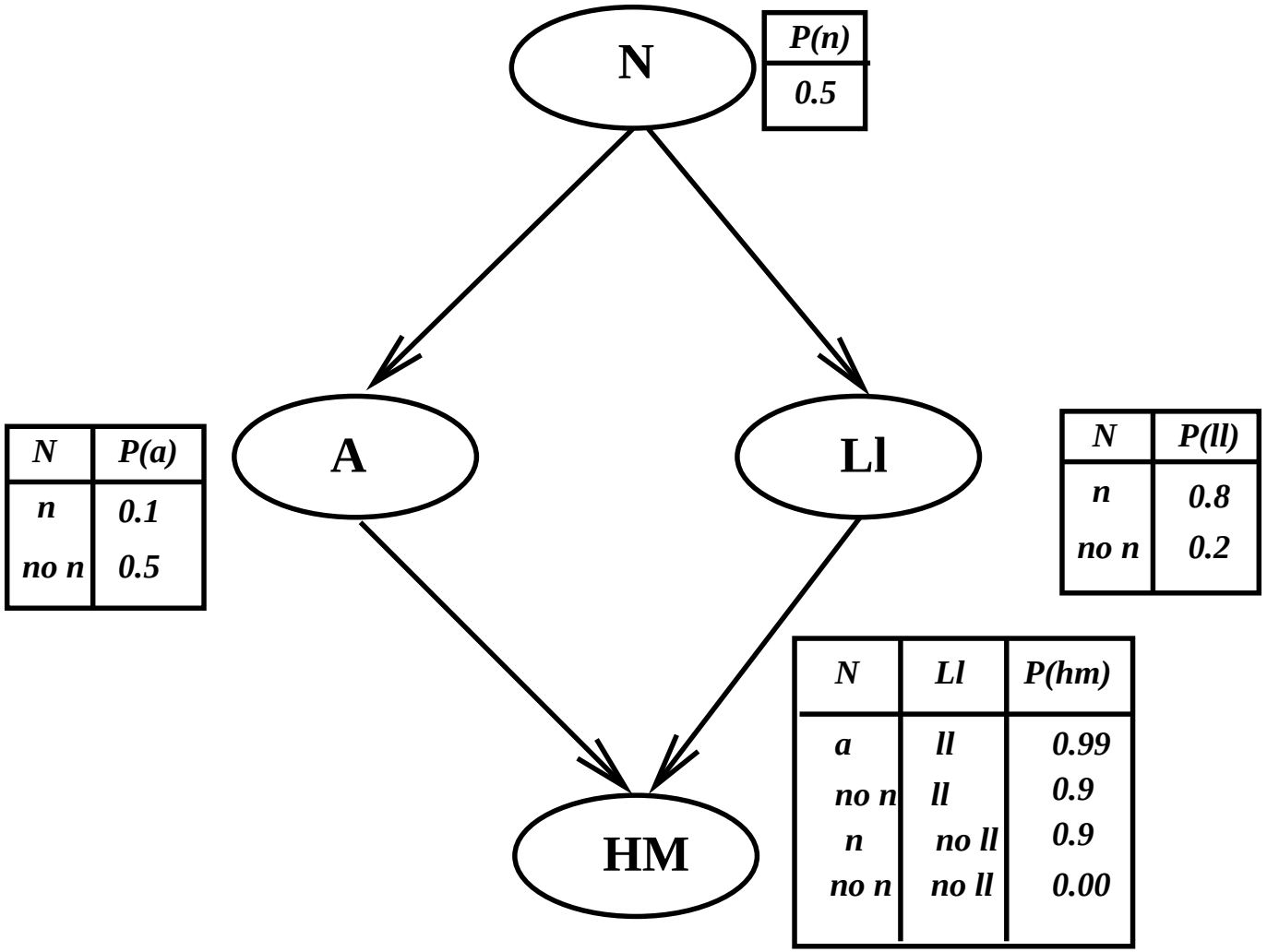
Muestreo

- Por *muestreo* (en ingles, *sampling*) respecto de una distribución de probabilidad entendemos métodos de generación de eventos, de tal manera que la probabilidad de generación de un evento dado coincide con la que indica la distribución
- El muestreo más sencillo: consideremos una v.a. booleana A tal que $\mathbf{P}(A) = \langle \theta, 1 - \theta \rangle$
  - Basta con tener un método de generación aleatoria y uniforme de números $x \in [0, 1]$
  - Si se genera $x < \theta$, se devuelve a; en caso contrario $\neg a$
  - En el límite, el número de muestras generadas con valor a entre el número de muestras totales es θ
- De manera sencilla, se generaliza esta idea para diseñar un procedimiento de muestreo respecto de la DCC que representa una red bayesiana

Ejemplo de red bayesiana (Russell & Norvig)

- Consideremos las siguientes variables aleatorias booleanas:
  - N: el cielo está nublado
  - A: el aspersor se ha puesto en marcha
  - LL: ha llovido
  - HM: la hierba está mojada
- Las relaciones causales y el conocimiento probabilístico asociado están reflejadas en la siguiente red bayesiana
  - Nótese que es un ejemplo de red que no está simplemente conectada.

Otro ejemplo: red bayesiana



Ejemplo de muestreo *a priori*

- Veamos cómo generar un evento aleatorio completo a partir de la DCC que especifica la red anterior
  - Muestreo de $\mathbf{P}(N) = \langle 0{,}5; 0{,}5\rangle$; aplicando el método anterior obtenemos n
  - Muestreo de $\mathbf{P}(A|n) = \langle 0{,}1; 0{,}9\rangle$; obtenemos $\neg a$
  - Muestreo de $\mathbf{P}(LL|n) = \langle 0{,}8; 0{,}2\rangle$; obtenemos ll
  - Muestreo de $\mathbf{P}(HM|\neg a, ll) = \langle 0{,}9; 0{,}1\rangle$; obtenemos hm
- El evento generado por muestreo ha sido $\langle n, \neg a, ll, hm\rangle$
- La probabilidad de generar ese evento es $0{,}5 \times 0{,}9 \times 0{,}8 \times 0{,}9 = 0{,}324$, ya que cada muestreo individual se realiza independientemente

Algoritmo de muestreo *a priori*

- Supongamos dada una *RED* bayesiana con un conjunto de variables $X_1, \ldots, X_n$ ordenadas de manera consistente con el orden implícito en la red

Algoritmo de muestreo *a priori*

FUNCION MUESTREO-A-PRIORI(RED)
1. PARA $i = 1, \ldots, n$ **HACER** x_i **el resultado de un muestreo de** $P(X_i | padres(X_i))$
2. Devolver $(x_1, \ldots, x_n)$

Propiedades del algoritmo de muestreo *a priori*

- Sea $S_{MP}(x_1, \ldots, x_n)$ la probabilidad de que el evento $(x_1, \ldots, x_n)$ sea generado por **`MUESTREO-A-PRIORI`**
- Es fácil ver (deduciéndolo *exclusivamente* del algoritmo) que $S_{MP}(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | padres(X_i))$
- Por tanto, si repetimos el muestreo anterior N veces y llamamos $N_{MP}(x_1, \ldots, x_n)$ al número de veces que se devuelve el evento $(x_1, \ldots, x_n)$, entonces

$$\frac{N_{MP}(x_1, \ldots, x_n)}{N} \approx P(x_1, \ldots, x_n)$$

  - Donde $\approx$ significa que en el límite (cuando N tiende a ∞) esa igualdad se convierte en exacta (según la ley de los grandes números)
  - Es lo que se llama una *estimación consistente*

Muestreo con rechazo

- La propiedad anterior será la base para los algoritmos de inferencia no exacta que veremos a continuación
  - Recordar que el problema de la inferencia probabilística es el de calcular $\mathbf{P}(X|\mathbf{e})$ donde $\mathbf{e}$ denota un conjunto de valores concretos observados para algunas variables y X es la variable de consulta
- Muestreo con rechazo: generar muestras y calcular la proporción de eventos generados en los que $\mathbf{e}$ "se cumple"
- Ejemplo: para estimar $\mathbf{P}(LL|a)$ generar 100 muestras con el algoritmo anterior; si de estas muestras hay 27 que tienen valor de A igual a a, y de estas 27, LL es ll en 8 de los casos y es $\neg ll$ en 19, entonces la estimación es:

$$\mathbf{P}(LL|a) \approx \mathit{Normaliza}(\langle 8; 19\rangle) = \langle 0{,}296; 0{,}704\rangle$$

- La respuesta exacta sería $\langle 0{,}3; 0{,}7\rangle$

Algoritmo de inferencia: muestreo con rechazo

- Entrada: una v.a. X de consulta, un conjunto de valores observados $\mathbf{e}$ para la variables de evidencia, una *RED* bayesiana (con n variables) y número N de muestras totales a generar

Algoritmo de muestreo con rechazo

FUNCION MUESTREO-CON-RECHAZO(X,e,RED,N)
1. Sea $N[X]$ **un vector con una componente por cada posible valor de la variable de consulta** X**, inicialmente todas a 0**
2. PARA $k = 1, \ldots, N$ **HACER**
2.1 Sea $(y1, \ldots, yn)$ **igual a MUESTREO-A-PRIORI(RED)**
2.2 SI $y = (y1, \ldots, yn)$ **es consistente con** e **entonces HACER**
$N[x]$ **igual a** $N[x] + 1$**, donde en el evento** y **la v.a.** X **toma el valor** x
3. Devolver NORMALIZA($N[X]$**)**

- A partir de las propiedades de **MUESTREO-A-PRIORI** se deduce que este algoritmo devuelve una estimación consistente de $\mathbf{P}(X|\mathbf{e})$
- Problema: se rechazan demasiadas muestras (sobre todo si el número de variables de evidencia es grande)

Ponderación por verosimilitud

- Es posible diseñar un algoritmo que sólo genere muestras consistentes con la observación $\mathbf{e}$
- Los valores de las variables de evidencia no se generan: quedan fijados de antemano
- Pero no todos los eventos generados "pesan" lo mismo: aquellos en los que la evidencia es más improbable deben contar menos
- Por tanto, cada evento generado va a ir a compañado de un peso igual al producto de las probabilidades condicionadas de cada valor que aparezca en $\mathbf{e}$

Ejemplo de ponderación por verosimilitud

- Supongamos que queremos calcular $\mathbf{P}(LL|a, hm)$; para generar cada muestra con su correspondiente peso w, hacemos lo siguiente ($w = 1{,}0$ inicialmente):
  - Muestreo de $\mathbf{P}(N) = \langle 0{,}5; 0{,}5\rangle$; obtenemos n
  - Como A es una variable de evidencia (cuyo valor es a) hacemos w igual a $w \times P(a|n)$ (es decir, $w = 0{,}1$)
  - Muestreo de $\mathbf{P}(LL|n) = \langle 0{,}8; 0{,}2\rangle$; obtenemos ll
  - HM es una variable de evidencia (con valor hm); por tanto, hacemos w igual a $w \times P(hm|a, ll)$ (es decir, $w = 0{,}099$)
- Por tanto, el muestreo devolvería $\langle n, a, ll, hm\rangle$ con un peso igual a 0,099
- Este evento contaría para $LL = true$, pero ponderado con 0,099 (intuitivamente, refleja que es poco probable que funcionen los aspersores un día de lluvia)

Algoritmo de inferencia: ponderación por verosimilitud

- Entrada: una v.a. X de consulta, un conjunto de valores observados $\mathbf{e}$ para la variables de evidencia, una *RED* bayesiana (con n variables) y un número N de muestras totales a generar

Algoritmo de ponderación por verosimilitud

FUNCION PONDERACION-POR-VEROSIMILITUD(X,e,RED,N)
1. Sea $W[X]$ **un vector con una componente para cada posible valor de la variable de consulta** X**, inicialmente todas a 0**
2. PARA $k = 1, \ldots, N$ **HACER**
2.1 Sea $[(y1, \ldots, yn), w]$ **igual a MUESTRA-PONDERADA(RED,e)**
2.2 Hacer $W[x]$ **igual a** $W[x] + w$**, donde en el evento** y **la v.a.** X **toma el valor** x
3. Devolver NORMALIZA($W[X]$**)**

Algoritmo de inferencia: ponderación por verosimilitud

Obtenición de muestras ponderadas

FUNCION MUESTRA-PONDERADA(RED,e)
1. Hacer $w = 1{,}0$
2. PARA $i = 1, \ldots, n$ **HACER**
2.1 SI la variable X_i **tiene valor** x_i **en** e **ENTONCES**
$w = w \times p(X_i = x_i | padres(X_i))$
2.2 SI NO, sea x_i **el resultado de un muestreo de**
$P(X_i | padres(X_i))$
3. Devolver $[(x_1, \ldots, x_n), w]$

- Se puede demostrar que el algoritmo **PONDERACION-POR-VEROSIMILITUD** devuelve una estimación consistente de la probabilidad buscada
- En el caso de que haya muchas variables de evidencia, el algoritmo podría degradarse, ya que la mayoría de las muestras tendrían un peso infinitesimal

Más sobre algoritmos de inferencia aproximada

- Existen muchos otros algoritmos de inferencia aproximada en redes bayesianas, más sofisticados que los vistos aquí
- Uno de ellos es el algoritmo Monte Carlo de cadenas de Markov
  - En él, cada evento se genera a partir del anterior, haciendo cambios aleatorios en el valor de las variables no observadas y dejando fijo el valor de las variables de evidencia

Sección 3

Sección 3
El teorema de Bayes

Formulación de la regla de Bayes

- De $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$ podemos deducir la siguiente fórmula, conocida como *regla de Bayes*

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

- Regla de Bayes para variables aleatorias:

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

  - recuérdese que esta notación representa un conjunto de ecuaciones, una para cada valor específico de las variables

- Versión con normalización:

$$\mathbf{P}(Y|X) = \alpha \cdot \mathbf{P}(X|Y)\mathbf{P}(Y)$$

- Generalización, en presencia de un conjunto $\mathbf{e}$ de observaciones:

$$\mathbf{P}(Y|X, \mathbf{e}) = \alpha \cdot \mathbf{P}(X|Y, \mathbf{e})\mathbf{P}(Y|\mathbf{e})$$

Ejemplo de uso de la regla de Bayes

- Sabemos que la probabilidad de que un paciente de meningitis tenga el cuello hinchado es 0.5 (relación causal)
- También sabemos la probabilidad (incondicional) de tener meningitis ($\frac{1}{50000}$) y de tener el cuello hinchado (0.05)
- Estas probabilidades provienen del conocimiento y la experiencia
- La regla de Bayes nos permite diagnosticar la probabilidad de tener meningitis una vez que se ha observado que el paciente tiene el cuello hinchado

$$P(m|h) = \frac{P(h|m)P(m)}{P(h)} = \frac{0{,}5 \times \frac{1}{50000}}{0{,}05} = 0{,}0002$$

- Alternativamente, podríamos haberlo hecho normalizando
  - $\mathbf{P}(M|h) = \alpha\langle P(h|m)P(m); P(h|\neg m)P(\neg m)\rangle$
  - Respecto de lo anterior, esto evita $P(h)$ pero obliga a saber $P(h|\neg m)$

Relaciones causa-efecto y la regla de Bayes

- Modelando probabilísticamente una relación entre una *Causa* y un *Efecto*:
  - La regla de Bayes nos da una manera de obtener la probabilidad de *Causa*, dado que se ha observado *Efecto*:

$$\mathbf{P}(Causa|Efecto) = \alpha \cdot \mathbf{P}(Efecto|Causa)\mathbf{P}(Causa)$$

- Nos permite *diagnosticar* en función de nuestro conocimiento de relaciones *causales* y de probabilidades *a priori*
- ¿Por qué calcular el diagnóstico en función del conocimiento causal y no al revés?
  - Porque es más fácil y robusto disponer de probabilidades causales que de probabilidades de diagnóstico (lo que se suele conocer es $\mathbf{P}(Efecto|Causa)$ y el valor que ha tomado *Efecto*).

Regla de Bayes: ejemplo

- Consideremos la siguiente información sobre el cáncer de mama
  - Un 1 % de las mujeres de más de 40 años que se hacen un chequeo tienen cáncer de mama
  - Un 80 % de las que tienen cáncer de mama se detectan con una mamografía
  - El 9.6 % de las que no tienen cáncer de mama, al realizarse una mamografía se le diagnostica cáncer erróneamente

Regla de Bayes: ejemplo

- Pregunta 1: ¿cuál es la probabilidad de tener cáncer si la mamografía así lo diagnostica?
  - Variables aleatorias: C (tener cáncer de mama) y M (mamografía positiva)
  - $\mathbf{P}(C|m) = \alpha\mathbf{P}(C, m) = \alpha\mathbf{P}(m|C)\mathbf{P}(C) = \alpha\langle P(m|c)P(c); P(m|\neg c)P(\neg c)\rangle = \alpha\langle 0{,}8 \cdot 0{,}01; 0{,}096 \cdot 0{,}99\rangle = \alpha\langle 0{,}008; 0{,}09504\rangle = \langle 0{,}0776; 0{,}9223\rangle$
  - Luego el 7.8 % de las mujeres diagnosticadas positivamente con mamografía tendrán realmente cáncer de mama

Reglas de Bayes: ejemplo

- Pregunta 2: ¿cuál es la probabilidad de tener cáncer si tras dos mamografías consecutivas en ambas se diagnostica cáncer?
  - Variables aleatorias: M_1 (primera mamografía positiva) y M_2 (segunda mamografía positiva)
  - Obviamente, no podemos asumir independencia incondicional entre M_1 y M_2
  - Pero es plausible asumir independencia condicional de M_1 y M_2 *dada* C
  - Por tanto, $\mathbf{P}(C|m_1, m_2) = \alpha\mathbf{P}(C, m_1, m_2) = \alpha\mathbf{P}(m_1, m_2|C)\mathbf{P}(C) = \alpha\mathbf{P}(m_1|C)\mathbf{P}(m_2|C)\mathbf{P}(C) = \alpha\langle P(m_1|c)P(m_2|c)P(c); P(m_2|\neg c)P(m_2|\neg c)P(\neg c)\rangle = \alpha\langle 0{,}8 \cdot 0{,}8 \cdot 0{,}01; 0{,}096 \cdot 0{,}096 \cdot 0{,}99\rangle = \langle 0{,}412; 0{,}588\rangle$
  - Luego aproximadamente el 41 % de las mujeres doblemente diagnosticadas positivamente con mamografía tendrán realmente cáncer de mama

Sección 6

Sección 6

Clasificación mediante modelos probabilísticos: naive Bayes

Clasificadores *Naive Bayes*

- Supongamos un conjunto de atributos $A_1, \ldots, A_n$ cuyos valores determinan un valor en un conjunto finito V de posibles "clases" o "categorías"
- Tenemos un conjunto de entrenamiento D con una serie de tuplas de valores concretos para los atributos, junto con su clasificación
- Queremos aprender un clasificador tal que clasifique nuevas instancias $\langle a_1, \ldots, a_n \rangle$
  - Es decir, el mismo problema en el tema de aprendizaje de árboles de decisión y de reglas (pero ahora lo abordaremos desde una perspectiva probabilística).

Clasificadores *Naive Bayes*

- Podemos diseñar un modelo probabilístico para un problema de clasificación de este tipo, tomando los atributos y la clasificación como variables aleatorias
- El valor de clasificación asignado a una nueva instancia $\langle a_1, \ldots, a_n \rangle$, notado v_{MAP} vendrá dado por

$$\underset{v_j \in V}{argmax}\, P(v_j | a_1, \ldots, a_n)$$

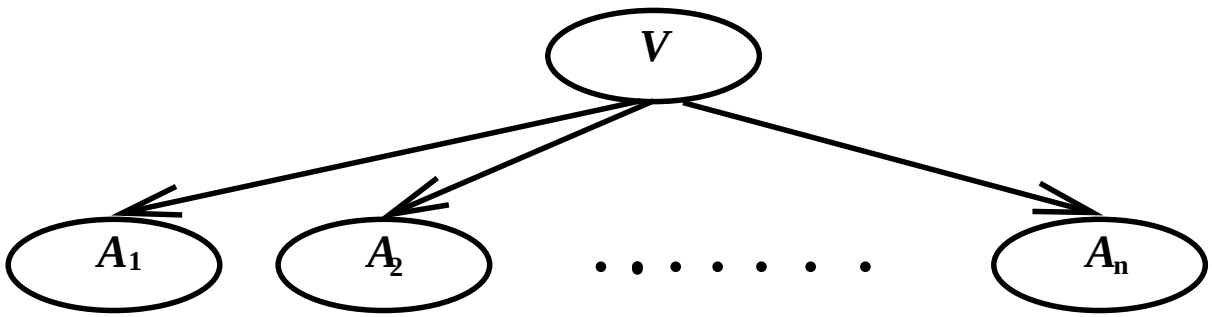
- Aplicando el teorema de Bayes podemos escribir

$$v_{MAP} = \underset{v_j \in V}{argmax}\, P(a_1, \ldots, a_n | v_j) P(v_j)$$

- Y ahora, simplemente estimar las probabilidades de la fórmula anterior a partir del conjunto de entrenamiento
- Problema: necesitaríamos una gran cantidad de datos para estimar adecuadamente las probabilidades $P(a_1, \ldots, a_n | v_j)$

Clasificadores *Naive Bayes*

- Podemos simplificar el aprendizaje suponiendo que los atributos son (mútuamente) condicionalmente independientes dado el valor de clasificación (de ahí lo de "naive")
- La situación se representa entonces por la red:



- En ese caso, tomamos como valor de clasificación:

$$v_{NB} = \underset{v_j \in V}{argmax}\, P(v_j) \prod_i P(a_i|v_j)$$

Estimación de probabilidades *Naive Bayes*

- Para el proceso de aprendizaje, sólo tenemos que estimar las probabilidades $P(v_j)$ (probabilidades *a priori*) y $P(a_i|v_j)$ (probabilidades *condicionadas*). Son muchas menos que en el caso general.
- Mediante cálculo de sus frecuencias en el conjunto de entrenamiento, obtenemos estimaciones de *máxima verosimilitud* de esas probabilidades:

$$P(v_j) = \frac{n(V = v_j)}{N} \qquad P(a_i|v_j) = \frac{n(A_i = a_i, V = v_j)}{n(V = v_j)}$$

  donde N es el número total de ejemplos, $n(V = v_j)$ es el número de ejemplos clasificados como v_j y $n(A_i = a_i, V = v_j)$ es el número de ejemplos clasificados como v_j cuyo valor en el atributo A_i es a_i.
- A pesar de su aparente sencillez, los clasificadores Naive Bayes tienen un rendimiento comparable al de los árboles de decisión, las reglas o las redes neuronales

Clasificador Naive Bayes: un ejemplo

- Vamos a aplicar el clasificador a un ejemplo ya conocido, usado en el tema de árboles de decisión:

Ej.	Cielo	Temperatura	Humedad	Viento	JugarTenis
D_1	Soleado	Alta	Alta	Débil	-
D_2	Soleado	Alta	Alta	Fuerte	-
D_3	Nublado	Alta	Alta	Débil	+
D_4	Lluvia	Suave	Alta	Débil	+
D_5	Lluvia	Baja	Normal	Débil	+
D_6	Lluvia	Baja	Normal	Fuerte	-
D_7	Nublado	Baja	Normal	Fuerte	+
D_8	Soleado	Suave	Alta	Débil	-
D_9	Soleado	Baja	Normal	Débil	+
D_{10}	Lluvia	Suave	Normal	Débil	+
D_{11}	Soleado	Suave	Normal	Fuerte	+
D_{12}	Nublado	Suave	Alta	Fuerte	+
D_{13}	Nublado	Alta	Normal	Débil	+
D_{14}	Lluvia	Suave	Alta	Fuerte	-

Clasificador Naive Bayes: un ejemplo

- Supongamos que queremos predecir si un día soleado, de temperatura suave, humedad alta y viento fuerte es bueno para jugar al tenis
- Según el clasificador Naive Bayes:

$$v_{NB} = \underset{v_j \in \{+,-\}}{argmax}\ P(v_j)P(soleado|v_j)P(suave|v_j)P(alta|v_j)P(fuerte|v_j)$$

- Así que necesitamos estimar todas estas probabilidades, lo que hacemos simplemente calculando frecuencias en la tabla anterior:
  - $p(+) = 9/14$, $p(-) = 5/14$, $p(soleado|+) = 2/9$, $p(soleado|-) = 3/5$, $p(suave|+) = 4/9$, $p(suave|-) = 2/5$, $p(alta|+) = 3/9$, $p(alta|-) = 4/5$, $p(fuerte|+) = 3/9$ y $p(fuerte|-) = 3/5$

Clasificador Naive Bayes: un ejemplo

- Por tanto, las dos probabilidades a posteriori son:
  - $P(+)P(\textit{soleado}|+)P(\textit{suave}|+)P(\textit{alta}|+)P(\textit{fuerte}|+) = 0{,}0070$
  - $P(-)P(\textit{soleado}|-)P(\textit{suave}|-)P(\textit{alta}|-)P(\textit{fuerte}|-) = 0{,}0411$
- Así que el clasificador devuelve la clasificación con mayor probabilidad a posteriori, en este caso la respuesta es $-$ (no es un día bueno para jugar al tenis)

Detalles técnicos sobre las estimaciones: log-probabilidades

- Tal y como estamos calculando las estimaciones, existe el riesgo de que algunas de ellas sean excesivamente bajas
- Si realmente alguna de las probabilidades es baja y tenemos pocos ejemplos en el conjunto de entrenamiento, lo más seguro es que la estimación de esa probabilidad sea 0
- Esto plantea dos problemas:
  - La inexactitud de la propia estimación
  - Afecta enormemente a la clasificación que se calcule, ya que se multiplican las probabilidades estimadas y por tanto si una de ellas es 0, anula a las demás
- Una primera mejora técnica, intentado evitar productos muy bajos: usar logaritmos de las probabilidades.
  - Los productos se transforman en sumas

$$v_{NB} = \underset{v_j \in V}{argmax}[log(P(v_j)) + \sum_i log(P(a_i|v_j))]$$

Detalles técnicos sobre las estimaciones: suavizado

- Problema en la estimaciones:
  - Probabilidades nulas o muy bajas, por ausencia en el conjunto de entrenamiento de algunos valores de atributos en algunas categorías
  - Sobreajuste
- Idea: *suponer* que tenemos ejemplos adicionales, cuyos valores se distribuyen teóricamente *a priori* de alguna manera.

Suavizado aditivo (o de Laplace)

- Un caso particular de lo anterior se suele usar para la estimación de las probabilidades condicionales en Naive Bayes:

$$P(a_i|v_j) = \frac{n(A_i = a_i, V = v_j) + k}{n(V = v_j) + k|A_i|}$$

  donde k es un número fijado y $|A_i|$ es el número de posibles valores del atributo $|A_i|$.
- Intuitivamente: se supone que además de los del conjunto de entrenamiento, hay k ejemplos en la clase v_j por cada posible valor del atributo A_i
- Usualmente $k = 1$, pero podrían tomarse otros valores
  - Elección de k: experimentando con los dstintos rendimientos sobre un *conjunto de validación*

Aplicaciones de las redes bayesianas

- Aplicaciones en empresas
  - Microsoft: *Answer Wizard (Office)*, diagnóstico de problemas de impresora, *TrueSkill* rankings y emparejamiento de jugadores en *Xbox Live*.
  - Intel: Diagnóstico de fallos de procesadores
  - Nasa: Ayuda a la decisión de misiones espaciales
- Otras aplicaciones: diagnóstico médico, exploraciones petrolíferas, *e-learning*, bioinformática, clasificación de textos, procesamiento de imágenes, identificación basada en ADN...
- En general, sistemas expertos que manejan incertidumbre

Bibliografía

- Russell, S. y Norvig, P. *Artificial Intelligence (A Modern Approach), 3rd edition* (Prentice–Hall, 2010)
  - Cap. 13: "Quantifying Uncertainty"
  - Cap. 14: "Probabilistic Reasoning"
- Russell, S. y Norvig, P. *Inteligencia artificial (Un enfoque moderno), segunda edición* (Prentice–Hall Hispanoamericana, 2004)
  - Cap. 13: "Incertidumbre"
  - Cap. 14: "Razonamiento Probabilístico"