Sistemas Inteligentes

Redes Bayesianas

Tema 7

Objetivos

- Conocer el modelo de representación más extendido actualmente para el razonamiento con incertidumbre en IA: las redes de Bayes
- Saber diseñar sistemas sencillos de toma de decisiones con información incierta basados en redes de Bayes

Contenidos

- Introducción
- Repaso de teoría de la probabilidad
- Redes Bayesianas
  - Dependencia e Independencia de variables
  - Modelado
  - Motores de Inferencia
- Ejercicios

Introducción
Noción de incertidumbre

- En muchos dominios de interés para la IA es necesario trabajar con **incertidumbre**.
  - "*Falta de conocimiento seguro y claro de algo*". (R.A.E.)
- Algunos ejemplos de incertidumbre
  - Ignorancia
    - "El paciente tiene fiebre" (y no sabemos nada más)
  - Vaguedad
    - "Hace calor"
  - Imprecisión
    - "La tarea dura unas 3 o 4 horas"
  - Información contradictoria
    - "Nunca digas nunca"

Introducción
Algunos ejemplos de razonamiento con incertidumbre

- Un paciente presenta tos, fiebre y dificultad para respirar
  - ¿Cuál es la probabilidad de que tenga gripe?
  - ¿Cuál es la probabilidad de que tenga neumonía?
  - Le hacemos una radiografía y aparecen manchas en el pulmón
  - ¿Cuál es **ahora** la probabilidad de que tenga gripe? ¿y neumonía?
- El wifi no funciona en el móvil
  - Hay un 40% de posibilidades de que sea culpa del router, un 30% de que sea culpa del proveedor, un 20% de que sea culpa del móvil y un 10% de que sea por otras causas
  - Sin embargo, veo que el portátil, que está conectado al mismo router, sí que funciona. ¿Qué probabilidad hay ahora de que haya un problema con el móvil?

Introducción

Aspectos relevantes de los sistemas de razonamiento con incertidumbre

- Es el tipo de razonamiento que se hace en la mayoría de las situaciones reales
- El razonamiento suele ser no monótono
  - El grado de convicción sobre ciertos hechos puede variar a lo largo del proceso de razonamiento
- Se trata de utilizar el conocimiento disponible de la mejor manera posible
  - Si el conocimiento aumenta, debería mejorar la calidad de las inferencias
- Los modelos clásicos como la Lógica o las Reglas de Producción no sirven
  - Son necesarios nuevos modelos que permitan representar y razonar con conocimiento incierto, impreciso, vago, contradictorio, …

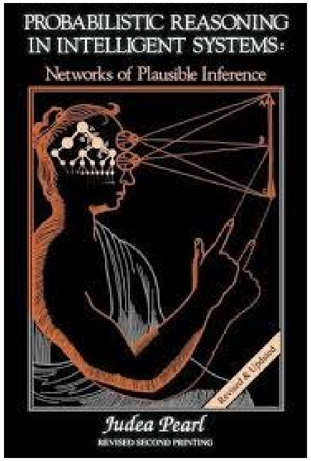
Introducción

Algunos modelos de razonamiento con incertidumbre

- Modelos simbólicos (extensiones de la lógica clásica)
  - Lógicas por Defecto (por defecto, algo es cierto)
  - Lógicas basadas en Modelos Mínimos
    - Hipótesis de Mundo Cerrado (CWA) (todo lo que no es cierto es falso)
    - Circunscripción (los modelos de los predicados son los de las fórmulas)
- Modelos numéricos
  - Modelos probabilistas
    - Teoría de Bayes (incertidumbre). REDES DE BAYES
    - Teoría de Dempster-Shafer (creencia, ignorancia)
  - Redes de neuronas
  - Lógica fuzzy o borrosa (vaguedad)
- Modelos mixtos
  - Sistemas de reglas con factores de certeza (MYCIN)

Introducción
Redes de Bayes

- Las redes de Bayes o redes bayesianas son consideradas como una de las contribuciones más importantes de la Inteligencia Artificial en las últimas décadas
- Fueron propuestas por Judea Pearl (premio Turing en 2012) en el año 1988



Repaso (I). Probabilidad de sucesos

- Resultados de un experimento
  - Espacio muestral U: conjunto de resultados posibles, sucesos elementales
  - Muestra o suceso aleatorio: subconjunto $A \subseteq U$ ($A=U$ suc. seguro, $A=\emptyset$ suc. imposible).
- Probabilidad de sucesos aleatorios
  - $P: \Sigma \subseteq 2^U \rightarrow \boldsymbol{R}$
    - $P(U) = 1$
    - $\forall A \subseteq U, P(A) \geq 0$
    - $\forall A, B \subseteq U, A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$
  - De estos "axiomas de la probabilidad" se deduce
    - $P(\emptyset) = 0$
    - $\forall A \subseteq U, P(A) \leq 1$
    - $\forall A \subseteq U, P(U \backslash A) = 1 - P(A)$
    - $\forall A, B \subseteq U, P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    - $\forall A_1, \cdots, A_n \subseteq U, A_1 \cap \ldots \cap A_n = \emptyset, P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n)$
- Interpretaciones de la probabilidad
  - Bayesiana: grado de creencia en que se va a producir un suceso
  - Frecuencia relativa con la que se produce un suceso
  - . . . .

Teoría de la Probabilidad
Repaso (II). Variable aleatoria. Distr. Probabilidad

- Variable aleatoria X (discreta) en un espacio (U, P)
  - $X: U \rightarrow \Omega_X$
    - Ω_X es el espacio de X
  - Ejemplos
    - $U = \{1,2,3,4,5,6\}$
    - $X \equiv$ Paridad: $\Omega_X=\{par, impar\}$
    - $Y \equiv$ Intervalo: $\Omega_Y=\{menor\_que\_tres, entre\_tres\_cuatro, mayor\_que\_tres\}$
- Función de distribución de probabilidad P de una variable aleatoria X
  - $P: \Omega_X \rightarrow \boldsymbol{R}$
    - $P(x_i) \equiv P(X = x_i) = P(\{e \in U | X(e) = x_i\})$
    - Notación: $P(X)$ es una distribución de probabilidad sobre X
  - Ejemplos
    - $\Omega_X: P(par) = P(\{2,4,6\}) = P(\{2\})+P(\{4\})+P(\{6\}) = 1/2$
    - $\Omega_Y: P(mayor\_que\_cuatro) = P(\{5,6\}) = 1/3$
- Informalmente: una variable aleatoria X es una variable que puede tomar valores en un dominio Ω_X. Cada valor x_i tiene una determinada probabilidad $P(x_i)$

Teoría de la Probabilidad
Repaso (III). Probabilidad conjunta y marginal

- Probabilidad conjunta P de un conjunto de variables $\boldsymbol{X} = \{X_1, \dots, X_n\}$
  - $P: \Omega_1 \times \dots \times \Omega_n \rightarrow [0,1]$
    - Notación: $P(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n)$
    - Debe cumplir: $\sum_{x_1,\dots,x_n} P(x_1, \dots, x_n) = 1$
    - Notación: $P(X_1, \dots, X_n)$ es la tabla de distribución conjunta
      - Si las variables son booleanas, entonces hay 2^n conjuntos de valores $x_1, \dots, x_n$ en la tabla

- Probabilidad marginal
  - $P(x_1, \dots, x_i) = \sum_{x_{i+1},\dots,x_n} P(x_1, \dots, x_i, x_{i+1}, \dots, x_n) = \sum_{x_{i+1}} P(x_1, \dots, x_i, x_{i+1})$
  - $P(x_1, \dots, x_i, x_{i+1}) = \sum_{x_{i+2},\dots,x_n} P(x_1, \dots, x_i, x_{i+1}, x_{i+2} \dots, x_n)$

- Ejemplo: Tenemos tres variables booleanas A, B, C

A	B	C	P(A,B,C)
F	F	F	0.1
F	F	V	0.2
F	V	F	0.05
F	V	V	0.05
V	F	F	0.3
V	F	V	0.1
V	V	F	0.05
V	V	V	0.15

- A partir de la distribución conjunta podemos calcular probabilidades marginales

$$P(a) = P(a, \neg b, \neg c) + P(a, \neg b, c) + P(a, b, \neg c) + P(a, b, c)$$

$$P(a, \neg b) = P(a, \neg b, c) + P(a, \neg b, \neg c) = 0.1 + 0.3$$

Teoría de la Probabilidad
Repaso (IV). Probabilidad condicionada

- Probabilidad condicionada de X dado un valor de Y
  - $P(x|y) \equiv P(X=x|Y=y) = \frac{P(x,y)}{P(y)}$, siempre que $P(y) \neq 0$
    - $P(X|y)$ es una distribución de probabilidad sobre X
    - $P(X|Y)$ es el conjunto de distribuciones $P(X|y)$ para todos los valores de Y
    - Se generaliza de forma natural para conjuntos de variables $\boldsymbol{X}$ e $\boldsymbol{Y}$, $P(\boldsymbol{X}|\boldsymbol{Y})$
- Ejemplo: tres variables booleanas A, B, C

A	B	C	P(A,B,C)
F	F	F	0.1
F	F	V	0.2
F	V	F	0.05
F	V	V	0.05
V	F	F	0.3
V	F	V	0.1
V	V	F	0.05
V	V	V	0.15

  - Las probs. condicionadas se calculan a partir de probs. marginales

$$P(b|c) = \frac{P(b,c)}{P(c)} = \frac{0.05+0.15}{0.2+0.05+0.1+0.15} = 0.40$$

$$P(a|\neg b, c) = \cdots$$

$$P(a, \neg b|c) = \cdots$$

- Ejemplo: *La probabilidad de tener gripe es de 1/40, la de tener dolor de cabeza es de 1/10 y la de tener ambas a la vez es de 1/80. ¿Cuál es la probabilidad de tener dolor de cabeza cuando tienes gripe? ¿Y la de tener gripe cuando tienes dolor de cabeza?*

Teoría de la Probabilidad
Repaso (V). Regla de la cadena. Regla de Bayes

- Regla de la cadena

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

- Regla de Bayes

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

  - Es la base de los motores de inferencia de los sistemas inteligentes que razonan bajo incertidumbre
  - $P(X)$ es la probabilidad de X "a priori"
  - $P(X|Y)$ es la probabilidad de X "a posteriori"

- Si X puede tomar los valores $\{x_1, \ldots, x_m\}$ se suele escribir como (marginalizando: $P(Y) = \sum_{i=1}^{m} P(Y|x_i)P(x_i)$)

$$P(X|Y) = \frac{P(Y|X)P(X)}{\sum_{i=1}^{m} P(Y|x_i)P(x_i)}$$

- Regla de la cadena generalizada

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

- Para un conjunto de variables $\{X_1, \dots, X_n\}$, la probabilidad de que tomen el valor $\{x_1, \dots, x_n\}$ se podría calcular como

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i | X_{i+1} = x_{i+1}, \dots, X_n = x_n)$$

Para lo cual tendríamos que conocer muchas dependencias condicionales

- Independencia de variables aleatorias
  - Si $P(X|Y) = P(X)$ para todos los valores de X e Y, entonces estas variables son (marginalmente) independientes. Es equivalente a
    - $P(Y|X) = P(Y)$
    - $P(X, Y) = P(X)P(Y)$

- **Consecuencia:** Si sabemos que hay independencia entre algunas variables (booleanas), no necesitamos conocer las 2^n entradas de la distribución conjunta

- **Ejemplo:** si una variable es independiente de las demás, se reduce un orden de magnitud
  - Si conocemos: $P(\neg C, D) = 0.1, P(C, \neg D) = 0.05, P(C, D) = 0.05, P(S) = 0.4$
  - Y que S es independiente de C y de D: $P(S|D) = P(S|C) = P(S) = 0.4$
  - Podemos calcular toda la distribución conjunta conociendo 4 valores

$$P(\neg C, \neg D) = 1 - (P(C, D) + P(\neg C, D) + P(C, \neg D)) = 0{,}8$$

$$P(\neg S) = 1 - P(S) = 0{,}6$$

$$P(\neg C, \neg D, \neg S) = P(\neg C, \neg D) \cdot P(\neg S) = 0{,}48$$

$$P(\neg C, \neg D, S) = P(\neg C, \neg D) \cdot P(S) = 0{,}32$$

C	D	S	Prob.
F	F	F	0,48
F	F	V	0,32
F	V	F	0,06
F	V	V	0,04
V	F	F	0,03
V	F	V	0,02
V	V	F	0,03
V	V	V	0,02

- Independencia de variables aleatorias
  - Si $P(X|Y) = P(X)$ para todos los valores de X e Y, entonces estas variables son (marginalmente) independientes. Es equivalente a
    - $P(Y|X) = P(Y)$
    - $P(X,Y) = P(X)P(Y)$
- La independencia de variables se puede establecer a partir del conocimiento del problema
- Ejemplo

  $C \equiv$ *Manuela tiene caries*

  $D \equiv$ *A Manuela le duele una muela*

  $S \equiv$ *Mañana hará sol*

  - Es bastante evidente que S es independiente de C y de D
  - Y que C y D no son independientes

Teoría de la Probabilidad
Repaso (VIII). Independencia condicional

- Independencia condicional
  - Si $P(X|Y,K) = P(X|K)$ para todos los valores de X, Y y K, entonces X e Y son independientes dado K. Es equivalente a
    - $P(Y|X,K) = P(Y|K)$
    - $P(X,Y|K) = P(X|K)P(Y|K)$
- Ejemplo

  $C \equiv$ *Manuela tiene caries*

  $M \equiv$ *A Manuela le duele una muela*

  $D \equiv$ *El dentista encuentra algo sospechoso*

  - Si no conocemos C, M y D son dependientes: $P(M|D) \neq P(M)$
  - Pero el hecho de que Manuela tiene caries explica M y D, luego: $P(M|C,D) = P(M|C)$
  - Y si Manuela no tiene caries entonces $P(M|\neg C,D) \neq P(M|\neg C)$

Teoría de la Probabilidad
Repaso (IX). Ejemplos de ind. cond.

- Otro ejemplo
  - Fumar provoca que los dientes amarilleen
  - Fumar causa cáncer de pulmón
  - En principio hay relación entre dientes amarillos y cáncer de pulmón
  - Cuando se conoce que la persona es fumadora, desaparece dicha relación

- Otro ejemplo más
  - La lluvia hace que el suelo se moje
  - El suelo mojado incrementa la posibilidad de resbalarse
  - El que haya lluvia o no influye sobre que una persona resbale
  - Cuando se sabe que el suelo está mojado, el que haya lluvia no influye

Teoría de la Probabilidad
Repaso (X). Dependencia condicional

- Dependencia condicional
  - Si $P(X,Y) = P(X)P(Y)$, es decir X e Y son independientes
  - Pero $P(X|Y,K) \neq P(X|K)$ entonces X e Y son dependientes dado K
  - A diferencia de la independencia, es suficiente con darse sólo para algunos valores de X,Y y K
- Ejemplo

  $L \equiv$ *Hoy llovió*

  $R \equiv$ *Hoy regaron*

  $M \equiv$ *El suelo está mojado*

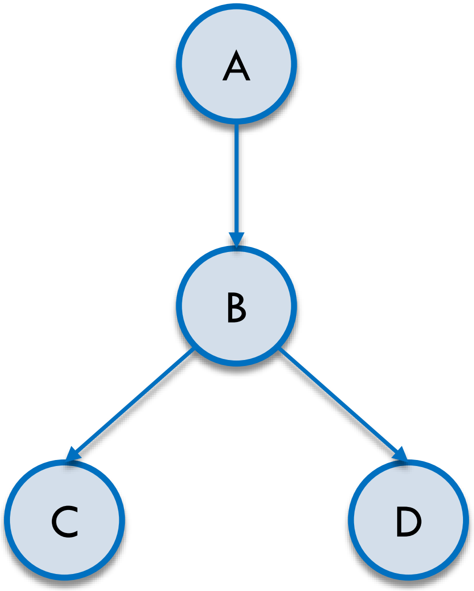
  - Si no conocemos M, L y R son independientes: $P(L,R) = P(L)P(R)$
  - Pero el hecho de que el suelo esté mojado hace que L y R sean dependientes, ya que si no es probable que regaran entonces es muy probable que lloviera, en este caso: $P(L|R,M) < P(L|R)$

Redes Bayesianas

- También se llaman redes causales o redes de creencia en la literatura
- Representan las relaciones de dependencia e independencia entre variables aleatorias
- **Muchos menos valores de probabilidad que las tablas de valores de distribución conjunta**
- **Mayor eficiencia para el cálculo de probabilidades conjuntas o marginales**
- Tienen muchas aplicaciones en las que el conocimiento que se maneja es incierto
  - Filtrado de correo
  - Reconocimiento de voz
  - Robótica
  - Diagnóstico médico
  - …

Redes Bayesianas
Estructura (DAG)

- Una red bayesiana es un modelo gráfico que representa relaciones de dependencia condicional entre un conjunto de variables aleatorias mediante un DAG (grafo dirigido acíclico)
  - Cada nodo del grafo es una variable aleatoria
  - Un arco de *X* a *Y* indica que estas variables son dependientes
    - *En un grafo causal: X es una causa de Y*
  - La ausencia de arcos entre dos nodos **no** significa independencia
    - *C* y *D* no son independientes, pero son condicionalmente independientes dado *B*

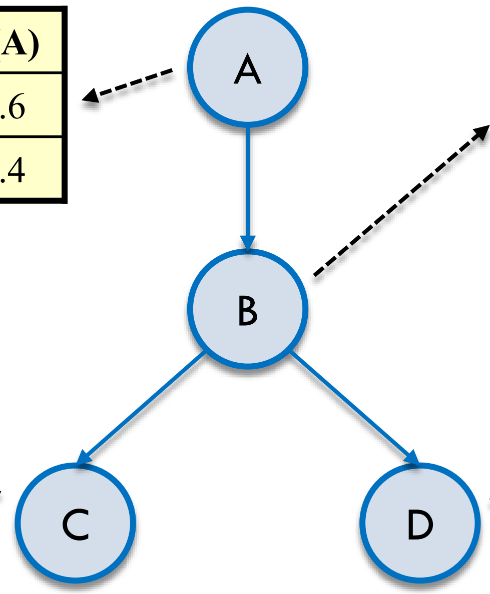


Redes Bayesianas
CPTs

- Cada nodo contiene una tabla de probabilidad condicional (CPT)
  - Expresa la distribución de probabilidad condicional $P(X_i|Padres(X_i))$ que cuantifica el efecto de los padres sobre el nodo

A	P(A)
F	0.6
V	0.4

A	B	P(B|A)
F	F	0.01
F	V	0.99
V	F	0.7
V	V	0.3



B	C	P(C|B)
F	F	0.4
F	V	0.6
V	F	0.9
V	V	0.1

B	D	P(D|B)
F	F	0.02
F	V	0.98
V	F	0.05
V	V	0.95

Redes Bayesianas

CPTs

- La suma de las probabilidades del nodo debe dar 1 para cada una de las combinaciones de valores de los padres
- Esto reduce **a la mitad** el número de valores que necesitamos conocer para cada tabla, si las variables son booleanas
- Si todas las variables son booleanas, para un nodo con *k* padres se requieren 2^k valores en cada nodo en lugar de 2^{k+1}

B	C	P(C|B)
F	F	0.4
F	V	0.6
V	F	0.9
V	V	0.1

1

1

B	C	P(C|B)
F	V	0.6
V	V	0.1

Redes Bayesianas
Cálculo de la probabilidad conjunta

- **Teorema de factorización**: Podemos calcular la distribución conjunta de todas las variables de la red $\{X_1, \dots X_n\}$ usando la fórmula:

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P\big(X_i = x_i | Padres(X_i)\big)$$

  - Donde $Padres(X_i)$ son los valores de los padres del nodo X_i

- Es menos costoso que aplicar la regla de la cadena

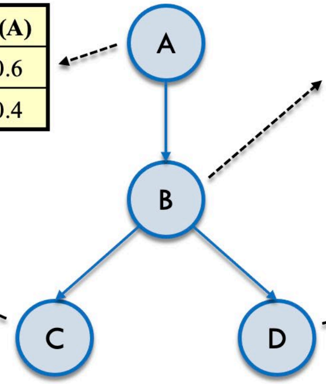
$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(X_i = x_i | X_{i+1} = x_{i+1}, \dots, X_n = x_n)$$

Redes Bayesianas

CPTs

A	P(A)
F	0.6
V	0.4

A	B	P(B|A)
F	F	0.01
F	V	0.99
V	F	0.7
V	V	0.3



B	C	P(C|B)
F	F	0.4
F	V	0.6
V	F	0.9
V	V	0.1

B	D	P(D|B)
F	F	0.02
F	V	0.98
V	F	0.05
V	V	0.95

- $P(C, \neg D, B, \neg A)$
- **Teorema de factorización**:

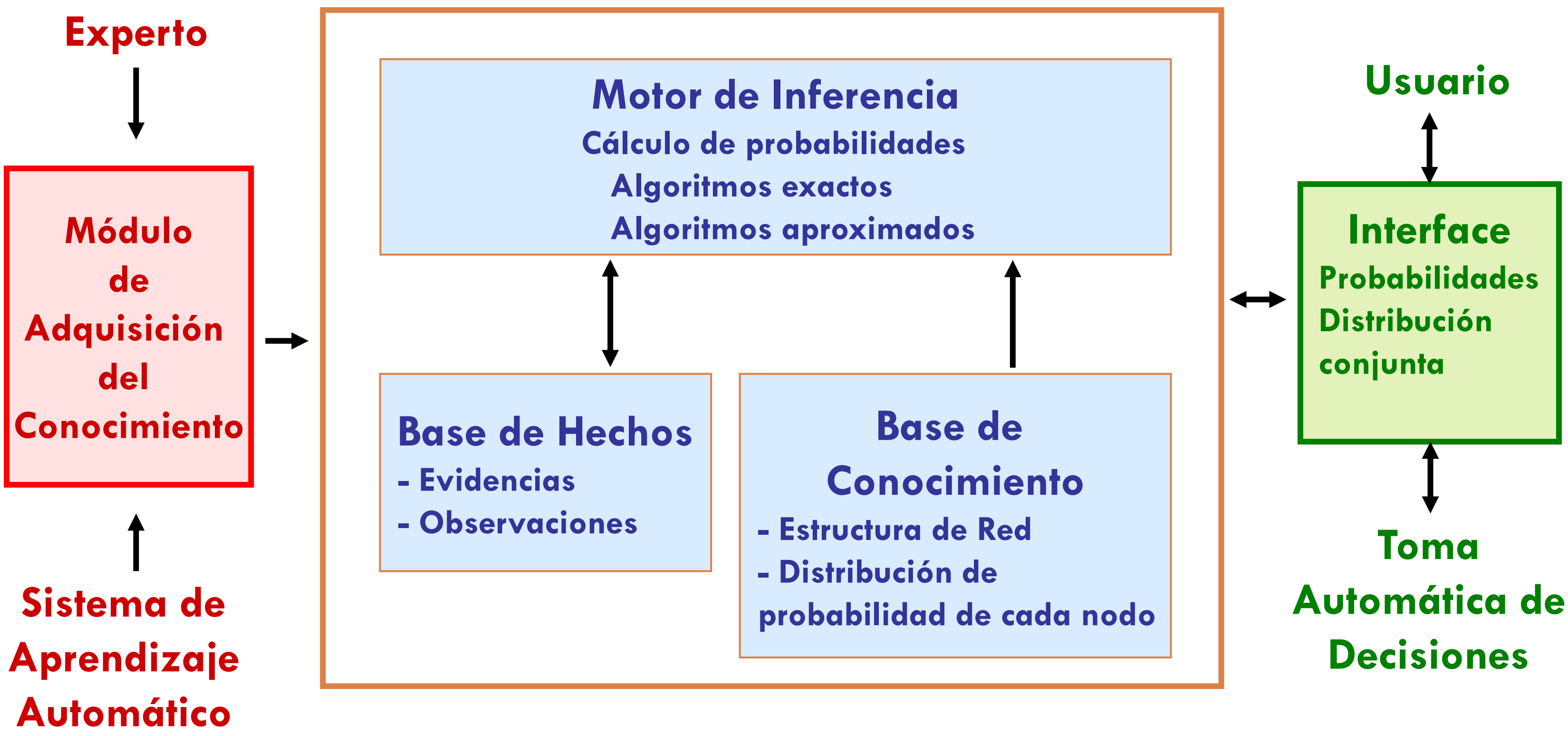
$$P(C, \neg D, B, \neg A)$$
$$= P(C|Padres(C))P(\neg D|Padres(D))P(B|Padres(B))P(\neg A|Padres(A))$$
$$= P(C|B)P(\neg D|B)P(B|\neg A)P(\neg A) = 0.1 \cdot 0.05 \cdot 0.99 \cdot 0.6$$

- Regla de la cadena

$$P(C, \neg D, B, \neg A)$$
$$= P(C|\neg D, B, \neg A)P(\neg D|B, \neg A)\ P(B|\neg A)\ P(\neg A)$$

Redes Bayesianas
SBC con RBs

- Redes Bayesianas para Representación de Conocimiento

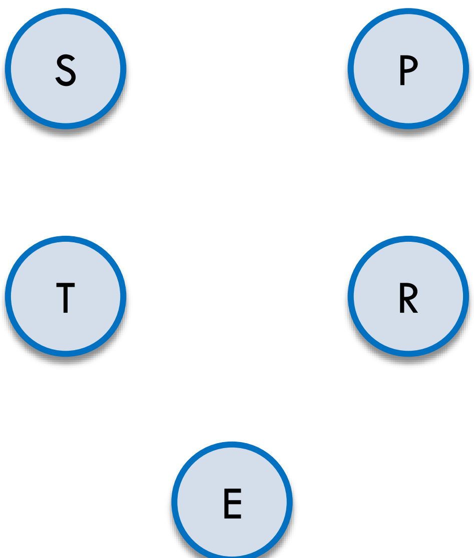


Construcción de una RB

- Ejemplo: Disponemos de un dominio con 5 posibles eventos:
  - P: Profesor que da la clase (Manuela o Andrew)
  - S: Hace sol
  - T: El profesor llega un poco tarde
  - R: La clase trata sobre robótica
  - E: La clase ha empezado a las 10:35
- Disponemos del siguiente conocimiento sobre el dominio:
  - El hecho de que haga sol no influye en quien da la clase.
  - Ambos profesores se retrasan con frecuencia debido al mal tiempo, pero Andrew es más tardón y por tanto es más probable que se retrase.
  - Es más probable que Andrew dé la clase sobre robótica.
  - Si el correspondiente profesor se retrasa, aumenta la probabilidad de que la clase no haya empezado a las 10:35.

Construcción de una RB
Un ejemplo (I)

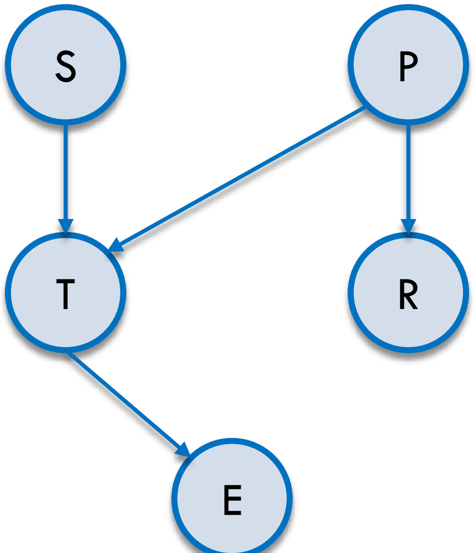
- Primer paso: Añade los nodos con las variables aleatorias
  - Cuidado: Elige las variables que quieres incluir. No todas tienen por qué ser aleatorias



P: Profesor que da la clase
S: Hace sol
T: El profesor llega un poco tarde
R: La clase trata sobre robótica
E: La clase ha empezado a las 10:35

Construcción de una RB
Un ejemplo (II)

- Segundo paso: Añade los arcos entre los nodos
  - La estructura debe ser acíclica
  - Si un nodo X tiene padres $\{Q_1, Q_2, \dots, Q_n\}$, cualquier variable que no sea un descendiente de X debe ser condicionalmente independiente de X dados $\{Q_1, Q_2, \dots, Q_n\}$



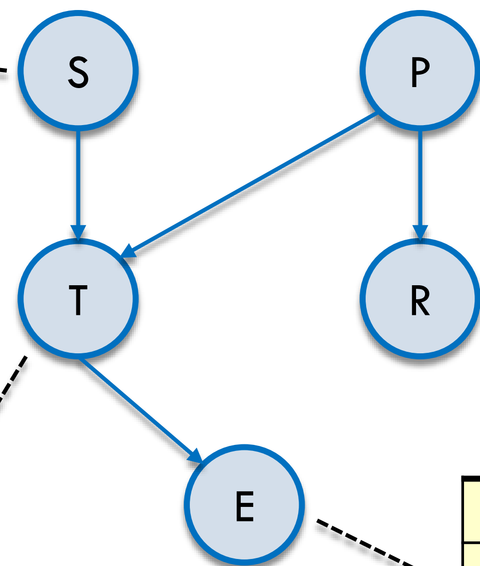
P: Profesor que da la clase
S: Hace sol
T: El profesor llega un poco tarde
R: La clase trata sobre robótica
E: La clase ha empezado a las 10:35

Construcción de una RB
Un ejemplo (III)

- Tercer paso: Añade los tablas de probabilidades
  - Una tabla por nodo con tantas filas como posibles combinaciones de valores tengan los padres

P(S)
0.3

P(P=Manuela)
0.60



P	P(R|P)
Manuela	0.30
Andrew	0.60

S	P	P(T|S,P)
F	Andrew	0.20
F	Manuela	0.10
V	Andrew	0.10
V	Manuela	0.05

T	P(E|T)
F	0.80
V	0.30

Construcción de una RB
Un ejemplo (IV)

- Necesitamos conocer **10 valores** para completar la red
  - Se puede calcular la distribución conjunta completa a partir de ellos

P(S)
0.30

P(P=Manuela)
0.60

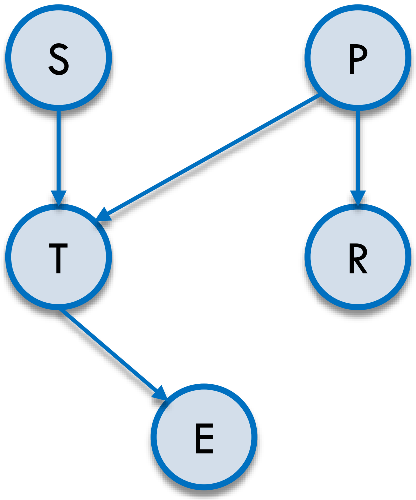
S	P	P(T|S,P)
F	Andrew	0.20
F	Manuela	0.10
V	Andrew	0.10
V	Manuela	0.05

P	P(R|P)
Manuela	0.30
Andrew	0.60

T	P(E|T)
F	0.80
V	0.30

Necesitaríamos conocer 2^5=**32 valores** si no utilizásemos Redes Bayesianas!!

Independencia en RB
Ejemplo



P: Profesor que da la clase
S: Hace sol
T: El profesor llega un poco tarde
R: La clase trata sobre robótica
E: La clase ha empezado a las 10:35

- Algunos ejemplos de dependencia/independencia condicional
  - *S* y *E* son condicionalmente independientes dado *T*
    - *Pero son dependientes si no conocemos T*
  - *T y R son condicionalmente independientes dado P*
    - *Pero son dependientes si no conocemos P*
  - *S y P son condicionalmente dependientes dados T o E*
    - *Concepto de "Eliminación de explicaciones"*

Criterio de *d*-separación
Definición.

Nodo convergentes

- Efecto común $X \rightarrow Z \leftarrow Y$ $$P(x,z,y) = P(x)P(y)P(z|x,y)$$

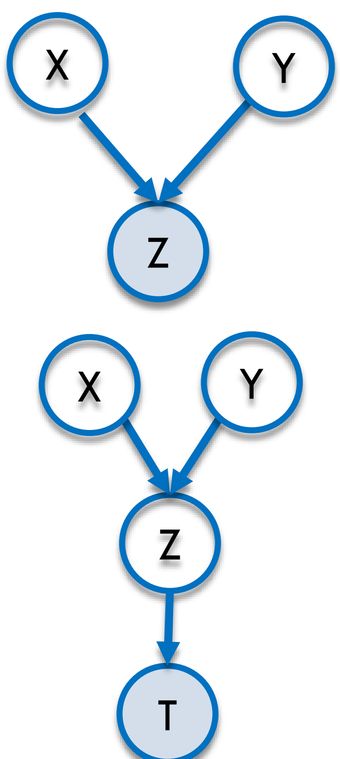
Nodos divergentes

- Causa común $X \leftarrow Z \rightarrow Y$ $$P(x,z,y) = P(z)P(x|z)P(y|z)$$

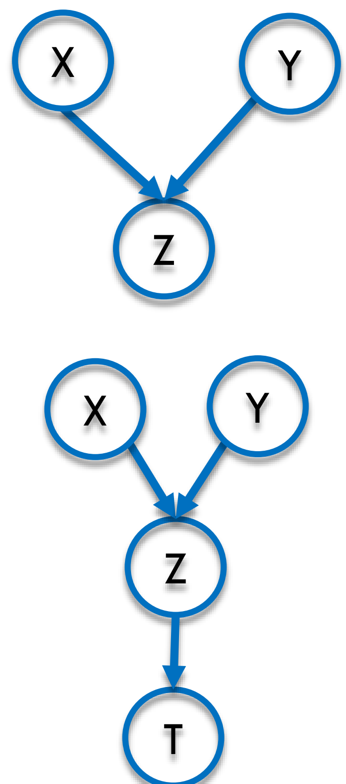
- Cadena causal $X \rightarrow Z \rightarrow Y$ $$P(x,z,y) = P(x)P(z|x)P(y|z)$$

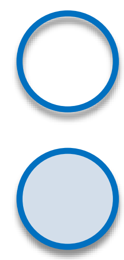
Criterio de *d*-separación
Definición.

- Efecto común
  - *X e Y condicionalmente dependientes*



  - *X e Y independientes*



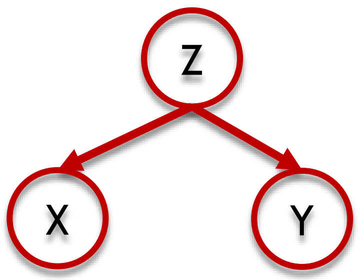


Variable NO observada

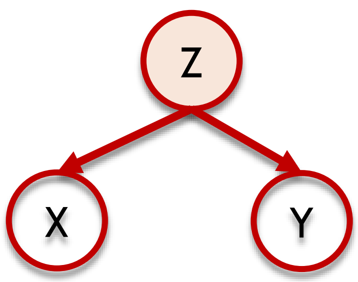
Variable observada

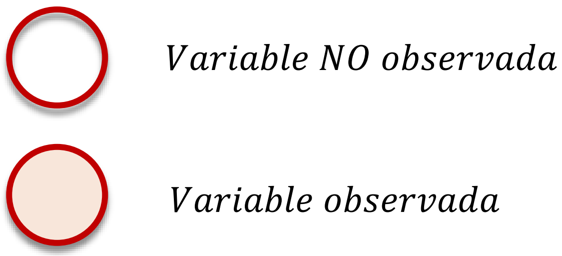
Criterio de *d*-separación
Definición.

- Causa común
  - *X* e *Y dependientes*



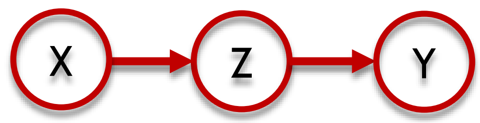
  - *X* e *Y condicionalmente independientes*



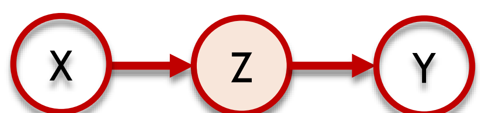


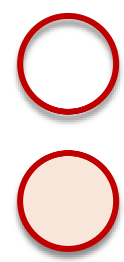
Criterio de *d*-separación
Definición.

- Cadena causal
  - *X e Y dependientes*



  - *X e Y condicionalmente independientes*





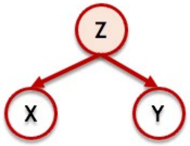
Variable NO observada

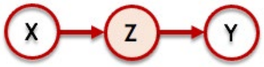
Variable observada

Criterio de *d*-separación

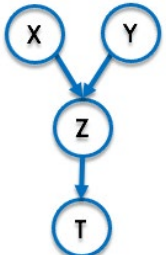
Definición. Camino bloqueado

- Un camino entre X e Y que pasa por Z está **bloqueado** …
  - Si el camino pasa por el nodo Z mediante una tripleta tipo causa común o cadena causal y el nodo Z está observado





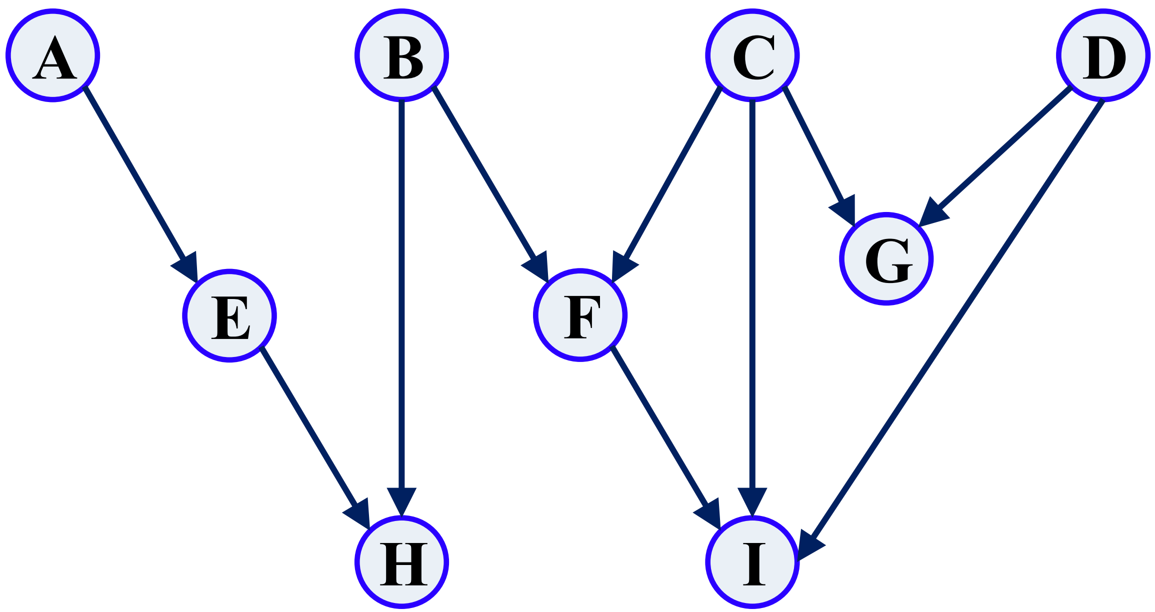
  - El camino pasa por el nodo Z mediante una tripleta efecto común y ni el nodo Z ni sus descendientes está observado.



- 2 variables X e Y son independientes si todos los caminos entre ambas están bloqueados. En otro caso X e Y son dependientes

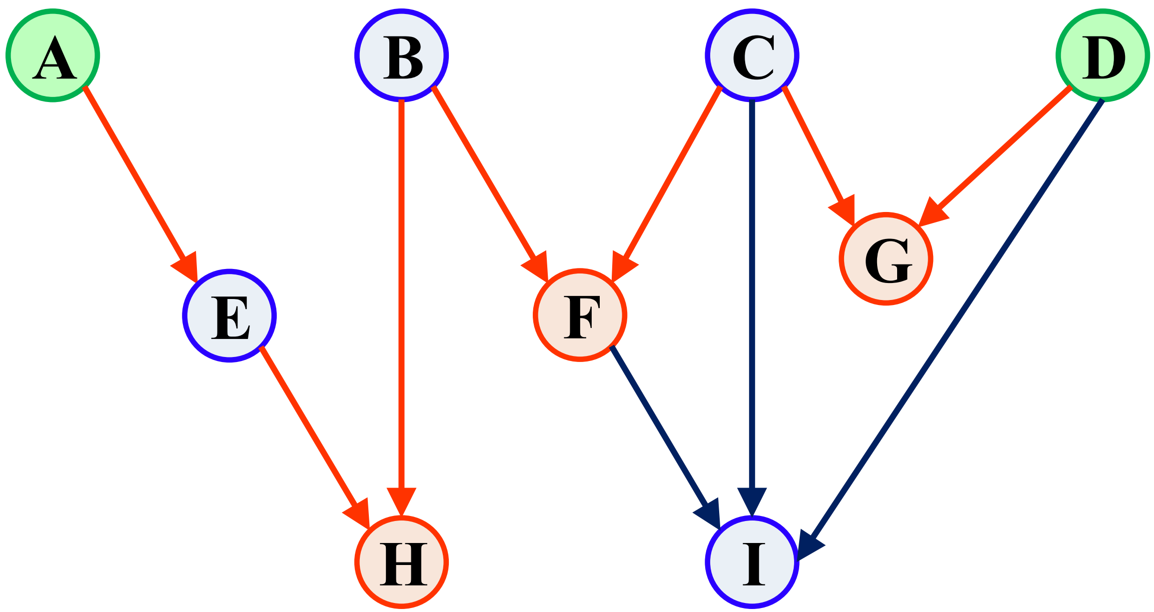
Criterio de *d*-separación
Ejemplo (I)

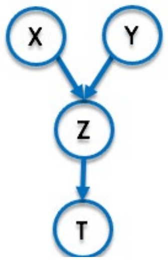
- ¿Son A y D independientes, si ningún nodo está observado?
  - Cuatro posibles caminos a comprobar



Criterio de *d*-separación
Ejemplo (II)

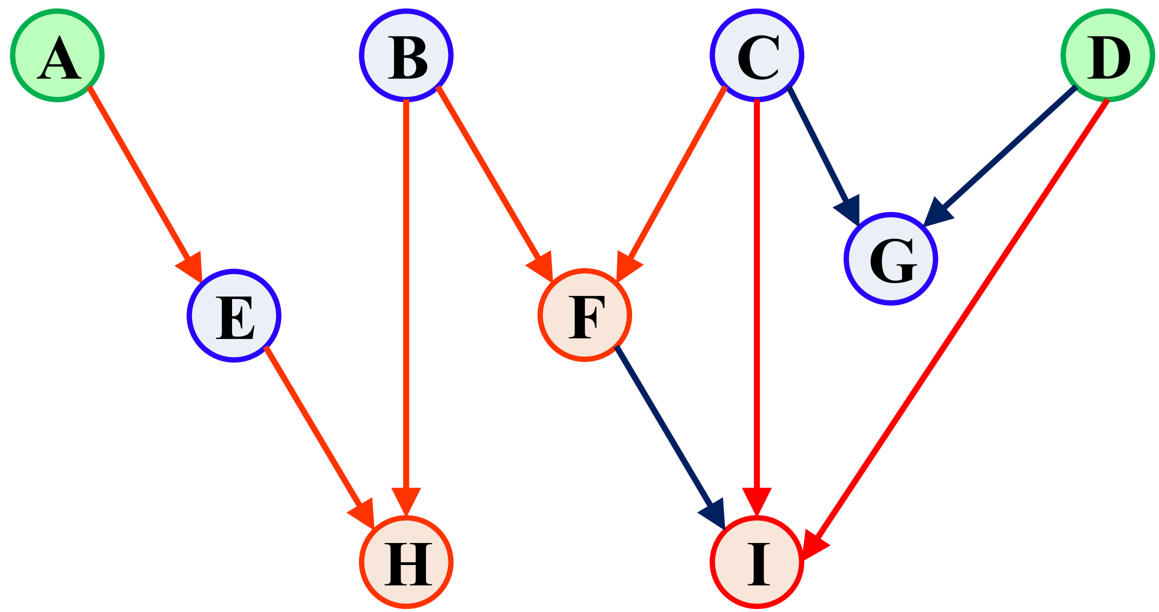
- ¿Son A y D independientes, si ningún nodo está observado?
  - Camino 1:

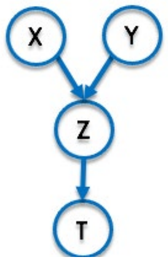




Criterio de *d*-separación
Ejemplo (III)

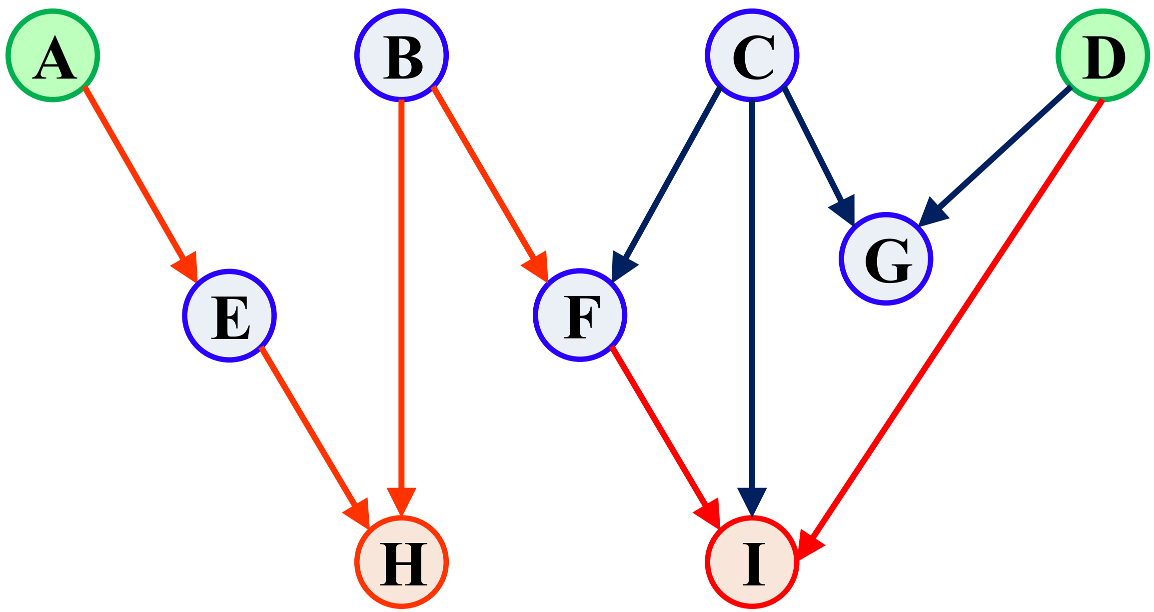
- ¿Son A y D independientes, si ningún nodo está observado?
  - Camino 2:

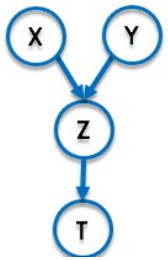




Criterio de *d*-separación
Ejemplo (III)

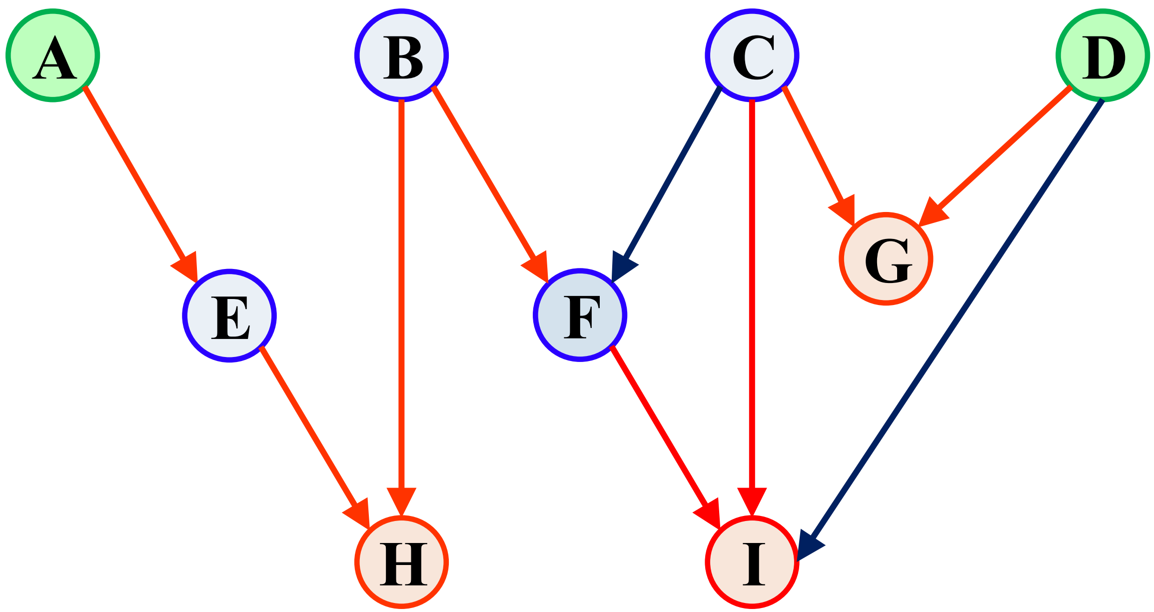
- ¿Son A y D independientes, si ningún nodo está observado?
  - Camino 3:

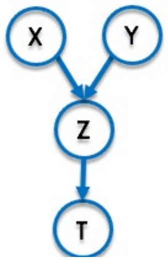




Criterio de d-separación
Ejemplo (IV)

- ¿Son A y D independientes, si ningún nodo está observado? **Sí, todos los caminos están bloqueados**
  - Camino 4:





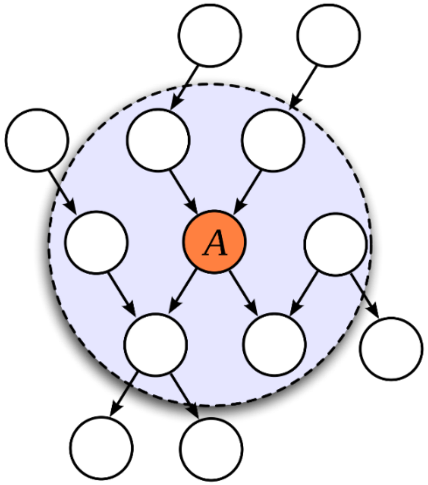
Independencia en RB

Markov Blanket o manto de Markov

- Dado un nodo A, su "Markov Blanket", $MB(A)$ viene dado por sus padres, sus hijos y el resto de los padres de sus hijos

- Dadas las variables de $MB(A)$, A es independiente de cualquier otra variable B de la red

$$P(A|MB(A),B) = P\left(A|MB(A)\right)$$

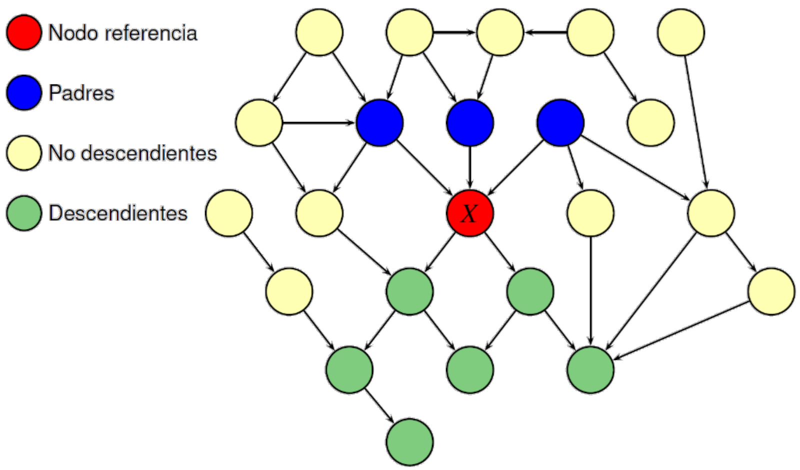


Independencia en RB
Condición de Markov

- **Condición de Markov:** un nodo X es independiente de todos sus no-descendientes, si conocemos el valor de todos sus padres (y solo de ellos)

$$P(X|Padres(X)) = P(X|Padres(X), NoDescendientes(X))$$

- Sólo se puede aplicar cuando los únicos nodos de la red de los que conocemos su valor son los padres del nodo X. No podemos conocer el valor de ningún nodo más



Inferencia en Redes Bayesianas

- La inferencia consiste en utilizar la red bayesiana para calcular la probabilidad de uno o varios eventos
- En general, en la inferencia se realizan preguntas como estas:
  - ¿Cuál es la probabilidad de que un (sub)conjunto de variables $\boldsymbol{X}$ tomen unos ciertos valores?

$$¿P(X_1 = x_1, \dots, X_n = x_n)?$$

  - ¿Cuál es la probabilidad de un conjunto de variables $\boldsymbol{X}$ si conocemos el valor de otro conjunto de variables $\boldsymbol{E}$?

$$¿P(\boldsymbol{X}|\boldsymbol{E})?$$

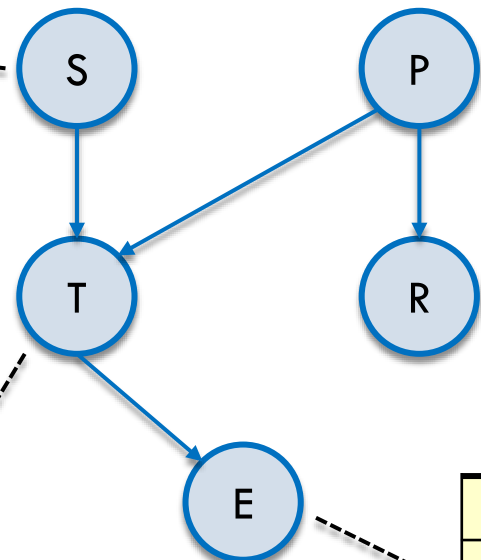
$\boldsymbol{X}$ = Variables "pregunta"
$\boldsymbol{E}$ = Variables "evidencia"

- Para realizar la inferencia exacta tenemos que utilizar la regla de Bayes, el teorema de probabilidad total, el cálculo de probabilidades marginales y el teorema de factorización

Inferencia exacta
Ejemplo 1

- ¿Cómo podemos calcular, por ejemplo, $P(E, \neg R, T, P = A, S)$?

$$P(E, \neg R, T, P = A, S) =$$
$$= P(E|T) \cdot P(\neg R|P = Andrew) \cdot P(T|S, P = Andrew) \cdot P(P = Andrew) \cdot P(S)$$



P(S)
0.3

P	P(P=x)
Manuela	0.6
Andrew	0.4

P	P(R|P)
Manuela	0.30
Andrew	0.60

P(~R/P=Andrew)=0.4

S	P	P(T|S,P)
F	Andrew	0.20
F	Manuela	0.10
V	Andrew	0.10
V	Manuela	0.05

T	P(E|T)
F	0.80
V	0.30

Inferencia exacta
Ejemplo 2

- ¿Cómo podemos calcular $P(R|E,\neg S)$?

$$P(R|E,\neg S) = \frac{P(R,E,\neg S)}{P(E,\neg S)} = \frac{P(R,E,\neg S)}{P(R,E,\neg S) + P(\neg R,E,\neg S)}$$

$$\begin{aligned} &P(R,E,\neg S) = \\ &= P(R,E,\neg S,T,P = Manuela) + P(R,E,\neg S,T,P = Andrew) \\ &+ P(R,E,\neg S,\neg T,P = Manuela) + P(R,E,\neg S,\neg T,P = Andrew) \end{aligned}$$

$$\begin{aligned} &P(\neg R,E,\neg S) = \\ &= P(\neg R,E,\neg S,T,P = Manuela) + P(\neg R,E,\neg S,T,P = Andrew) \\ &+ P(\neg R,E,\neg S,\neg T,P = Manuela) + P(\neg R,E,\neg S,\neg T,P = Andrew) \end{aligned}$$

- Cada una de esas probabilidades se calcularía como se hizo en el ejercicio anterior (teorema de factorización)

Inferencia exacta

Coste computacional

- La complejidad computacional crece exponencialmente con el número de variables aleatorias
  - No es factible en redes de gran tamaño

- La inferencia en redes bayesianas es un problema **NP-duro**

- Es común recurrir a técnicas *aproximadas* de inferencia
  - Mucho más rápidas
  - Permiten hacer inferencia en grandes redes
  - Dan resultados bastante aproximados al valor real

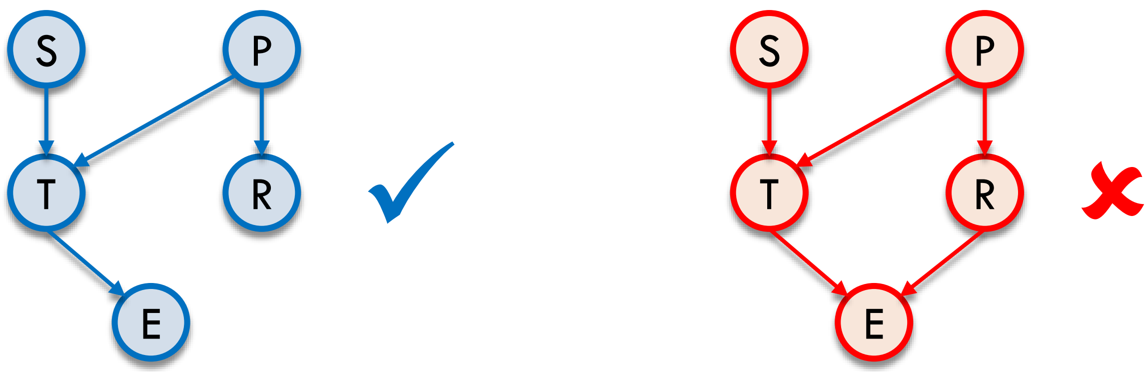
Algoritmos de inferencia en Redes Bayesianas

- Algoritmos exactos
  - Algoritmos para redes específicas
    - Árboles y Poliárboles (Kim y Pearl): Complejidad lineal
  - Algoritmos para redes generales
    - Eliminación de variables
    - Árbol de uniones (Lauritzen y Spiegelhalter)

- Algoritmos aproximados
  - Aproximar una distribución con una tolerancia dada es NP-duro.
  - Algunos algoritmos
    - Muestreo estocástico
    - Ponderación de la verosimilitud (Fung y Chang)
    - Markov Chain Monte Carlo

Poliárboles

- Un **poliárbol** es un grafo dirigido acíclico en el que, entre cada par de nodos, hay como máximo un camino



- Si una red es un árbol o un poliárbol, entonces existe un algoritmo de inferencia de complejidad lineal
- Algunos métodos funcionan convirtiendo una red a un árbol o un poliárbol
- El coste computacional de convertir una red en un árbol o un poliárbol puede ser muy elevado

Inferencia exacta
Eliminación de variables

- El algoritmo básico de inferencia repite ciertos cálculos varias veces
- **Idea:** Guardar los resultados de estos cálculos para evitar repetirlos.
- La complejidad sigue siendo exponencial en el número de variables de los factores intermedios.
- Es necesario buscar "buenos" órdenes de eliminación de variables para reducir el tamaño de los factores intermedios.
- Buscar una ordenación óptima tiene también complejidad exponencial.
  - Se emplean diversas heurísticas para encontrar la ordenación óptima.

Inferencia aproximada

Muestreo estocástico

- Consiste en generar un conjunto aleatorio de asignaciones para nuestras variables que, si es lo suficientemente grande, tendrá la misma probabilidad que la distribución de probabilidad conjunta.
- *Ejemplo*: Queremos calcular $P(E_1|E_2)$
  - Creamos N muestras aleatorias y contamos el número de ocurrencias de los siguientes casos:
    - N_C: Número de muestras en las que se cumple E_2
    - N_S: Número de muestras en las que se cumplen E_1 y E_2
  - Si N es lo suficientemente grande, entonces:
    - N_C/N es una buena estimación de $P(E_2)$
    - N_S/N es una buena estimación de $P(E_1, E_2)$
  - Por tanto, podemos estimar:

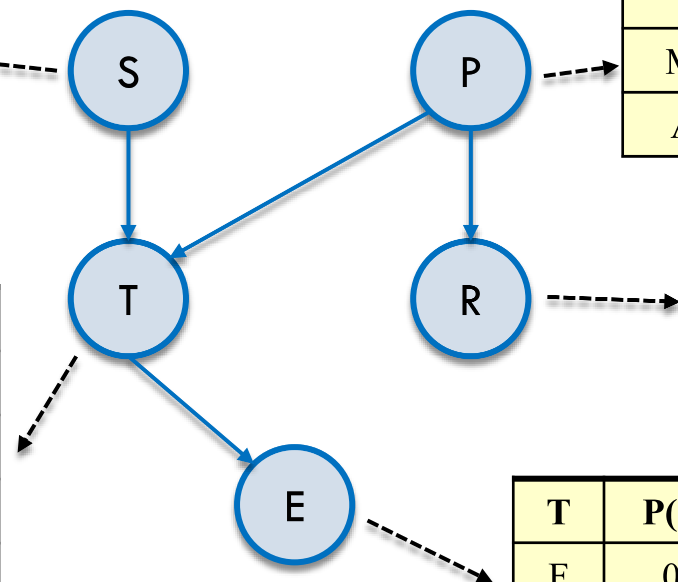
$$P(E_1|E_2) = P(E_1, E_2)/P(E_2) \approx N_S/N_C$$

Muestreo estocástico

- ¿Cómo generar cada muestra aleatoria?
  - Se ordenan las variables según un orden topológico de la red (de padres a hijos)
    - Llamaremos a las variables $X_1, \dots, X_n$, donde los padres de un nodo X_i deben ser un subconjunto de $\{X_1, \dots, X_{i-1}\}$
  - Desde $i = 1$ hasta $i = n$, hacer
    - Busca los padres $X_{p(i)}^1, \dots, X_{p(i)}^{n(i)}$ del nodo X_i
    - Basándote en los valores $x_{p(i)}^1, \dots, x_{p(i)}^{n(i)}$ asignados a los padres
      - Busca en la tabla de probabilidades: $P(X_i = x | X_{p(i)}^1 = x_{p(i)}^1, \dots, X_{p(i)}^{n(i)} = x_{p(i)}^{n(i)})$
      - Elige aleatoriamente $x_i = x$ según esa probabilidad
  - $\{x_1, x_2, \dots, x_n\}$ es la "muestra" de la distribución de probabilidad conjunta de $X_1, X_2, \dots, X_n$

Muestreo estocástico
Ejemplo

- Ejemplo: Generemos una muestra para calcular $P(R|E,\neg S)$



P(S)
0.3

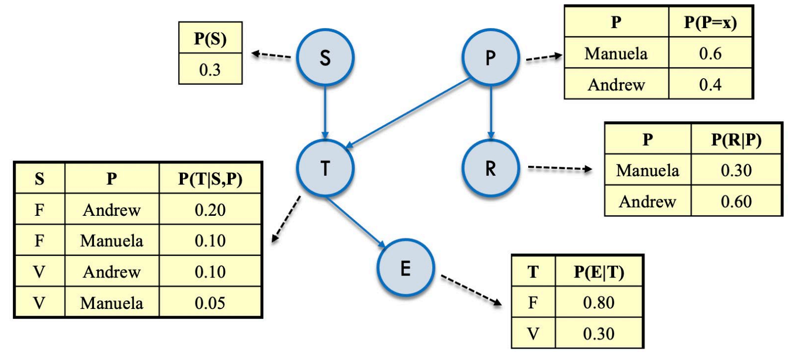
P	P(P=x)
Manuela	0.6
Andrew	0.4

P	P(R|P)
Manuela	0.30
Andrew	0.60

S	P	P(T|S,P)
F	Andrew	0.20
F	Manuela	0.10
V	Andrew	0.10
V	Manuela	0.05

T	P(E|T)
F	0.80
V	0.30

Muestreo estocástico
Ejemplo



P(S)
0.3

P	P(P=x)
Manuela	0.6
Andrew	0.4

P	P(R|P)
Manuela	0.30
Andrew	0.60

S	P	P(T|S,P)
F	Andrew	0.20
F	Manuela	0.10
V	Andrew	0.10
V	Manuela	0.05

T	P(E|T)
F	0.80
V	0.30

- Siguiendo nuestro ejemplo:
  1. Ordena las variables según un orden topológico
     - Por ejemplo: *S, P, T, R, E*
  2. Elige aleatoriamente *S*=V con probabilidad 0.3.
     - Por ejemplo, nos sale que *S*=V
  3. Elige aleatoriamente *P*=*Manuela* con probabilidad 0.6 y *P*=*Andrew* con probabilidad 0.4.
     - Por ejemplo, nos sale que *P*=*Andrew*
  4. La probabilidad de *T*=V depende de *S* y *P*.
     - Dado que *S*=V y *P*=*Andrew*, entonces *T*=V con prob. 0.1. Nos sale *T*=F
  5. La probabilidad de *R*=V depende de *P*.
     - Dado que *P*=*Andrew*, entonces *R*=V con prob. 0.6. Nos sale *R*=V
  6. La probabilidad de *E*=V depende de *T*.
     - Dado que *T*=F, entonces *E*=V con prob. 0.8. Supongamos que sale *E*=V
- La muestra generada {*S*,*P*=*Andrew*, ¬*T*,*R*,*E*} cumple E_1={*R*} pero no E_2={*E*, ¬*S*}, con lo que la muestra se descarta

Ponderación de la verosimilitud
Motivación

- Imagina que estamos a medio camino en la simulación:
  - En E_2 tenemos la restricción $X_i = x$
  - Estamos a punto de generar un valor aleatorio para X_i. Dados los valores asignados a sus padres, vemos que $P(X_i = x|Padres(X_i)) = p$
  - Generaremos un valor para $X_i = x$
    - Será $X_i = x$ con probabilidad p
    - O bien, $X_i \neq x$ con probabilidad (1- p) y esa simulación se descarta
- Problema del muestreo estocástico:
  - Cuando la probabilidad de que se cumplan todos los eventos de E_2 es muy baja, la mayoría de muestras son inútiles (no tienen ningún efecto en N_C y N_S)
- **Idea:** Generaremos siempre $X_i = x$, pero ponderaremos la muestra con el peso p para compensar

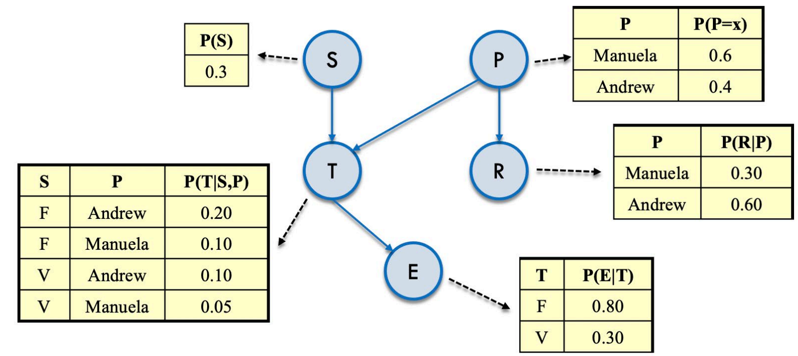
Ponderación de la verosimilitud

- Algoritmo para el cálculo aproximado de $P(E_1|E_2)$

Sea $X_1, \dots, X_n$ un orden topológico de la red
Inicializa $N_S = 0, N_C = 0$
Para $i = 1, \dots, N$ (N: número de muestras)
    Inicializa $w = 1$
    Para $j = 1, \dots, n$
        Si X_j es un evento de E_2 (cuyo valor es x)
            $X_j = x$ y $w = w \cdot P(X_j = x|Padres(X_j))$
        Si no
            Asigna un valor aleatorio a X_j dependiendo del valor de sus padres
    Fin Para
    $N_C = N_C + w$
    Si la muestra generada cumple E_1
        $N_S = N_S + w$
Fin Para
Devuelve N_S/N_C

Ponderación de la verosimilitud

- Siguiendo nuestro ejemplo para $P(R|E,\neg S)$:
  1. Orden topológico: *S, P, T, R, E*
  2. Inicializamos $w = 1$
  3. **Forzamos** S a Falso y actualizamos w
     - P(S) = 0.3, P(¬S)=0.7, luego $w = 0.7 \cdot w = 0.7$



P(S)
0.3

P	P(P=x)
Manuela	0.6
Andrew	0.4

P	P(R|P)
Manuela	0.30
Andrew	0.60

S	P	P(T|S,P)
F	Andrew	0.20
F	Manuela	0.10
V	Andrew	0.10
V	Manuela	0.05

T	P(E|T)
F	0.80
V	0.30

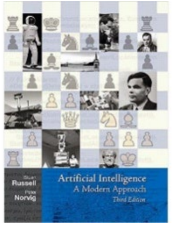
  4. Elige aleatoriamente *P=Manuela* con probabilidad 0.6 y *P=Andrew* con probabilidad 0.4.
     - Por ejemplo, nos sale que *P=Andrew*
  5. La probabilidad de *T*=V depende de *S* y *P*.
     - Dado que *S*=F y *P=Andrew*, entonces *T*=V con prob. 0.2. Nos sale *T*=F
  6. La probabilidad de *R*=V depende de *P*.
     - Dado que *P=Andrew*, entonces *R*=V con prob. 0.6. Nos sale *R*=V
  7. **Forzamos** el valor de E a V y actualizamos w
     - Dado que *T*=F, entonces *E*=V con prob. 0.8.
     - Por tanto: $w = 0.8 \cdot w = 0.8 \cdot 0.7 = 0.56$

Markov Chain Monte Carlo

- La ponderación de verosimilitud supone una mejora frente al muestreo estocástico
- Su rendimiento se sigue degradando si hay muchas variables evidencia
  - Unas pocas muestras tendrán casi la totalidad del peso
- Alternativa: Algoritmos de tipo Markov Chain Monte Carlo (MCMC)
  - No generan desde cero cada una de las muestras
  - Generan cada muestra aplicando un cambio aleatorio a la muestra anterior
- Algunos ejemplos:
  - *Gibbs sampling*
  - *Slice sampling*
  - *Langevin Markov chain Monte Carlo*

Bibliografía básica

- Section IV. Uncertain Knowledge and Reasoning



- Capítulo 6. Sistemas basados en modelos probabilísticos