Aprendizaje automatizado

Modelos gráficos probabilistas

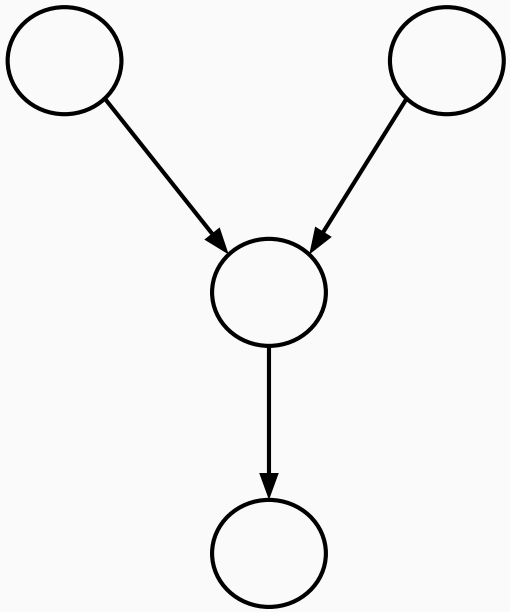
Gibran Fuentes-Pineda

Abril 2020

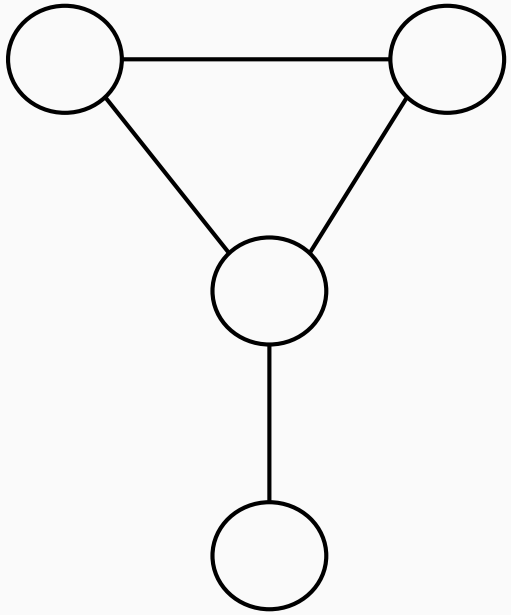
Modelos gráficos probabilistas

- Define una familia de distribuciones conjuntas de probabilidad sobre un conjunto de variables aleatorias

Grafo dirigido



Grafo no dirigido



Redes bayesianas

- **Nodos** representan variables aleatorias
- **Aristas** representan dependencias entre variables aleatorias
- La red representa relaciones causa-efecto
- Aprovecha independencias para especificar de forma compacta la distribución conjunta completa

Representación, inferencia y aprendizaje

- **Representación**: especificación de variables aleatorias y sus dependencias e independencias
- **Inferencia**: consulta de probabilidades en la red dado el modelo y ciertas evidencias
- **Aprendizaje**: obtener la topología y/o los parámetros de las distribuciones de la red a partir de ejemplos

Representación

- Estructura del grafo: variables aleatorias y dependencias e independencias

- Distribución conjunta se expresa por la distribución condicional de cada nodo dados sus padres (factorización)

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i | Padres(x_i))$$

Construyendo redes bayesianas

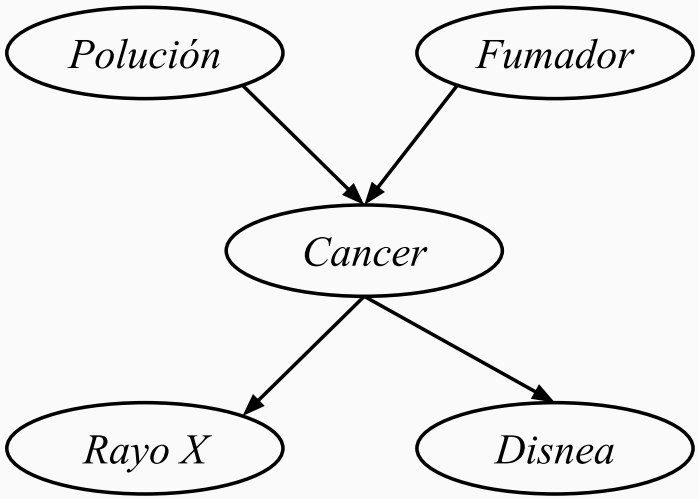
1. Elige el orden de las variables
2. Usa la regla del producto para obtener las conexiones
3. Reduce la complejidad aprovechando independencias

Construyendo redes bayesianas

1. Elige un orden de las variables
2. Usa la regla del producto para obtener las conexiones
3. Reduce la complejidad aprovechando independencias

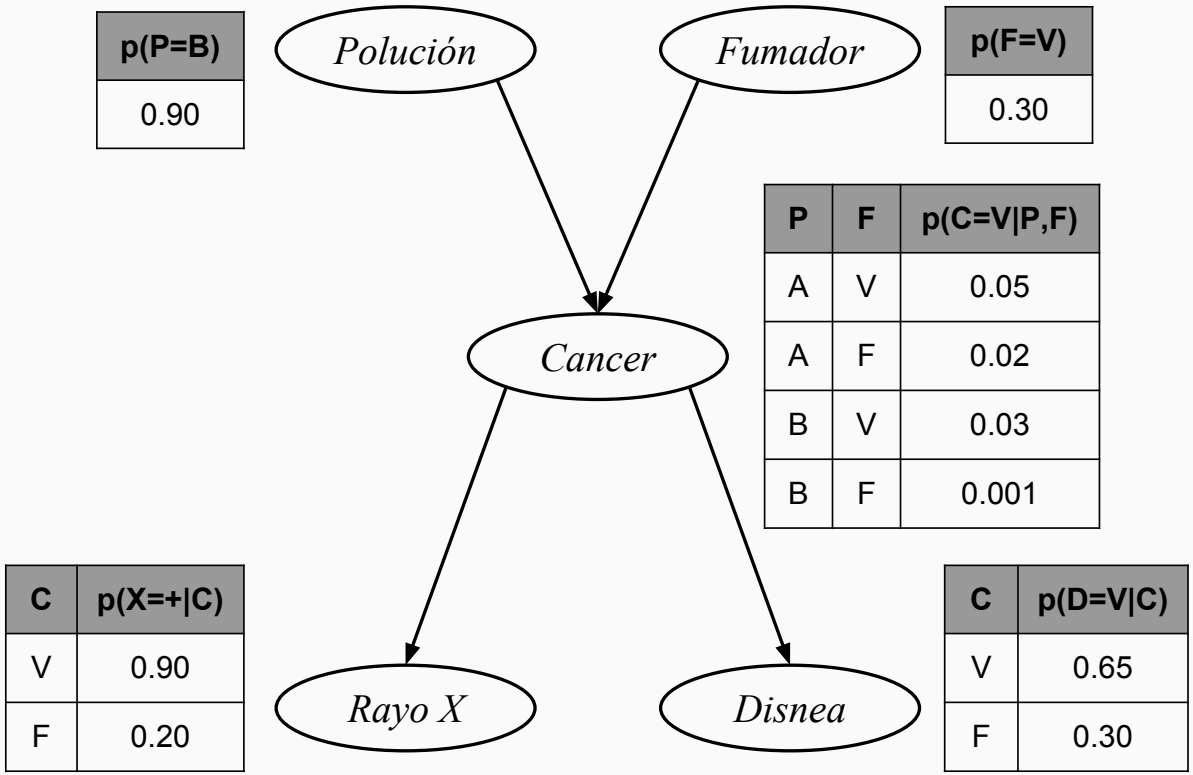
¡El orden importa!

Ejemplo: cáncer



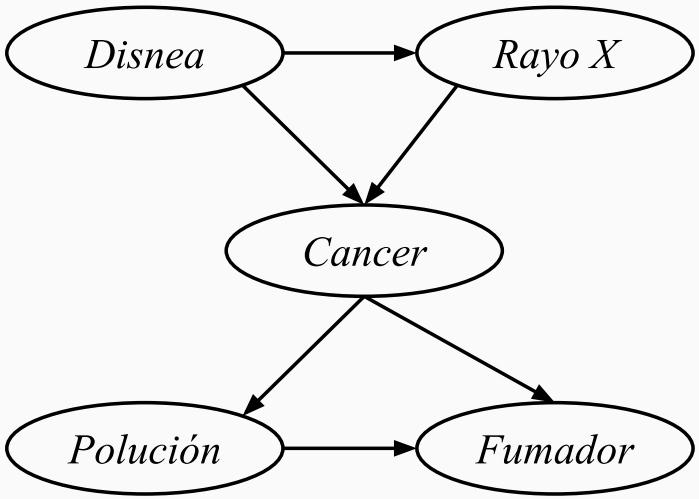
Ejemplo de BAI (Korb y Nicholson 2010)

Ejemplo: cáncer



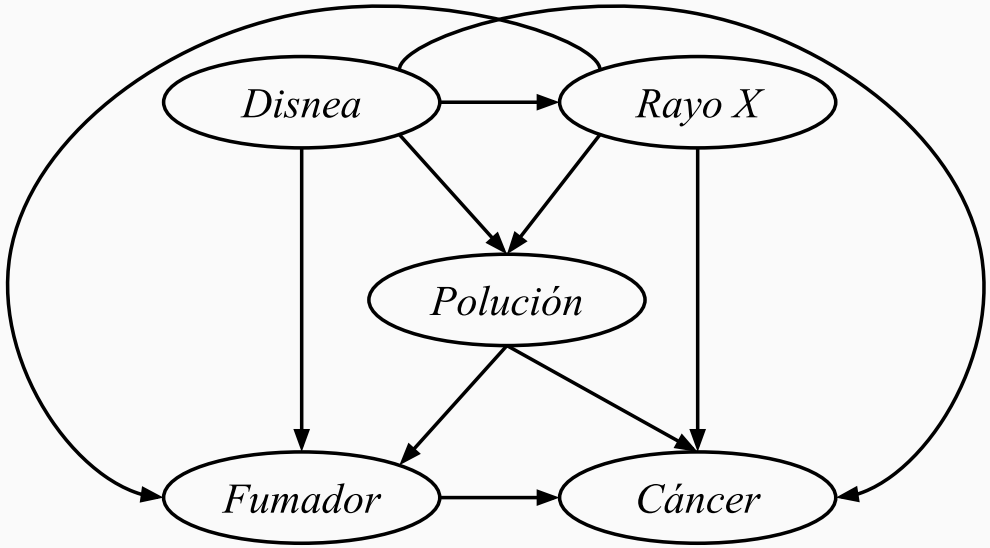
Ejemplo de BAI (Korb y Nicholson 2010)

Ejemplo: cáncer visto de otra manera



Ejemplo de BAI (Korb y Nicholson 2010)

Ejemplo: cáncer visto de otra manera



Ejemplo de BAI (Korb y Nicholson 2010)

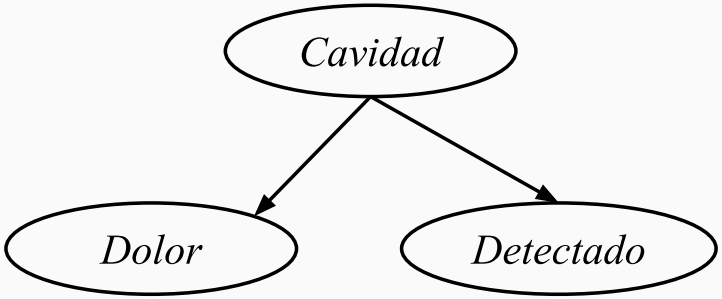
Mapa I y mapa D

- **Mapa-I**: No hay dependencias directas en la distribución conjunta que no estén especificadas en el grafo explícitamente
- **Mapa-D**: Todas las aristas en el grafo corresponden a dependencias directas en la distribución conjunta
- **Mapa perfecto**: Modelo gráfico con mapa-I y mapa-D

Distribución conjunta completa

- Contiene toda la información necesaria para obtener cualquier probabilidad.

- Por ejemplo



	$Do = \text{Sí}$		$Do = \text{No}$	
	$De = \text{Sí}$	$De = \text{No}$	$De = \text{Sí}$	$De = \text{No}$
$Ca = \text{Sí}$	0.108	0.012	0.072	0.008
$Ca = \text{No}$	0.016	0.064	0.144	0.576

Ejemplo de AIMA (Russel y Norvig 2009)

Cálculo de probabilidades marginales

- Para obtener una probabilidad marginal a partir de las probabilidades conjuntas aplicamos la regla de la suma.

- Por ejemplo

$$P(Ca = \text{Sí}) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

	$Do = \text{Sí}$		$Do = \text{No}$	
	$De = \text{Sí}$	$De = \text{No}$	$De = \text{Sí}$	$De = \text{No}$
$Ca = \text{Sí}$	0.108	0.012	0.072	0.008
$Ca = \text{No}$	0.016	0.064	0.144	0.576

Ejemplo de AIMA (Russel y Norvig 2009)

Cálculo de probabilidades condicionales

- Para obtener una probabilidad condicional a partir de la distribución conjunta aplicamos la regla del producto.

- Por ejemplo

$$P(Ca = \text{Sí} | Do = \text{Sí}) = \frac{P(Ca = \text{Sí}, Do = \text{Sí})}{P(Do = \text{Sí})}$$
$$= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.2$$

	$Do = \text{Sí}$		$Do = \text{No}$	
	$De = \text{Sí}$	$De = \text{No}$	$De = \text{Sí}$	$De = \text{No}$
$Ca = \text{Sí}$	0.108	0.012	0.072	0.008
$Ca = \text{No}$	0.016	0.064	0.144	0.576

Ejemplo de AIMA (Russel y Norvig 2009)

Número de parámetros para distribución conjunta

- ¿Cuántas probabilidades tenemos que calcular si agregamos la variable discreta Te con 10 posibles valores?

Número de parámetros para distribución conjunta

- ¿Cuántas probabilidades tenemos que calcular si agregamos la variable discreta Te con 10 posibles valores?

- Crecimiento exponencial: para n variables discretas de K valores sería K^n

Número de parámetros para distribución conjunta

- ¿Cuántas probabilidades tenemos que calcular si agregamos la variable discreta Te con 10 posibles valores?
- Crecimiento exponencial: para n variables discretas de K valores sería K^n
- Los modelos gráficos reducen la complejidad factorizando la distribución conjunta en distribuciones condicionales y aprovechando las relaciones de independencia

Aprovechando la independencia

- Las aristas corresponden a dependencias directas entre variables
- La ausencia de aristas captura las independencias absolutas y condicionales

Recordando la independencia condicional

- La variable aleatoria x es independiente de y dado z si

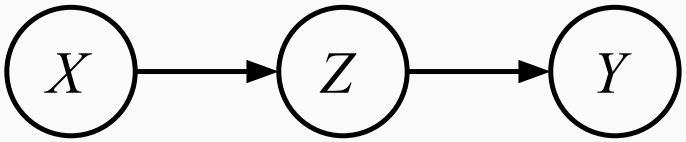
$$P(x, y|z) = P(x|z)P(y|z)$$
$$P(x|y, z) = P(x|z)$$

- La independencia condicional se denota con el símbolo $\perp\!\!\!\perp$

$$x \perp\!\!\!\perp y \mid z$$

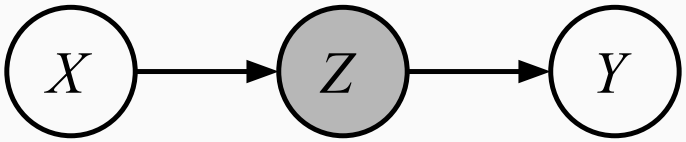
Independencia condicional: cadenas causales

$$x \not\perp\!\!\!\perp y \mid \varnothing$$



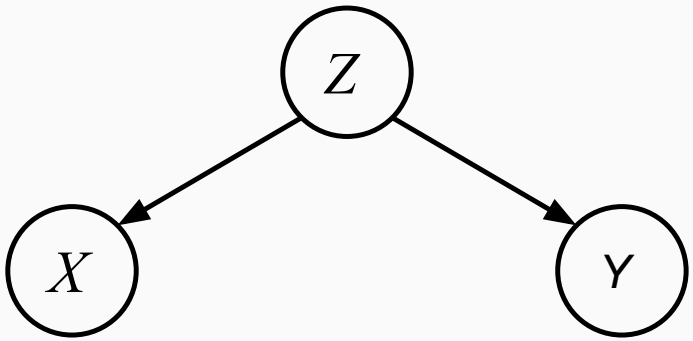
Independencia condicional: cadenas causales

$$x \perp\!\!\!\perp y \mid z$$



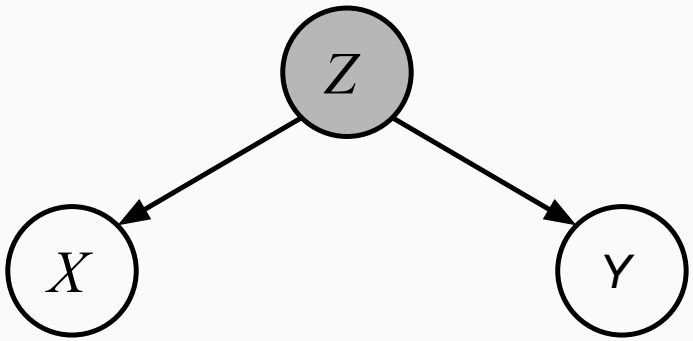
Independencia condicional: causas comúnes

$$X \not\perp\!\!\!\perp Y \mid \varnothing$$



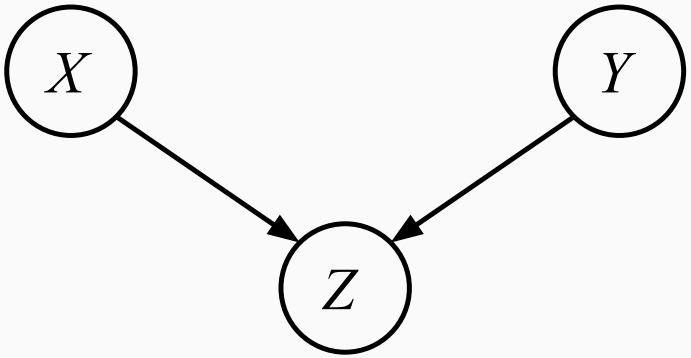
Independencia condicional: causas comúnes

$x \perp\!\!\!\perp y \mid z$



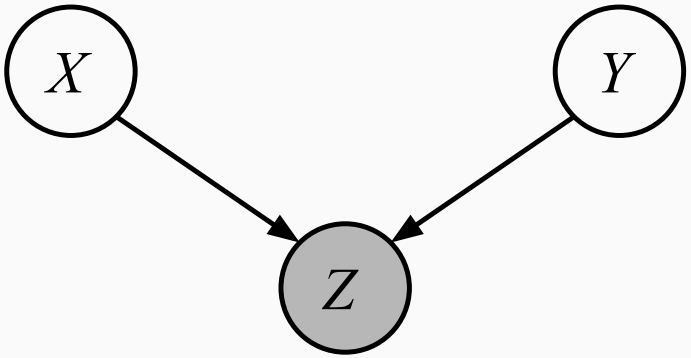
Dependencia condicional: efectos comúnes

$$x \perp\!\!\!\perp y \mid \varnothing$$



Dependencia condicional: efectos comúnes

$$x \not\perp\!\!\!\perp y \mid z$$

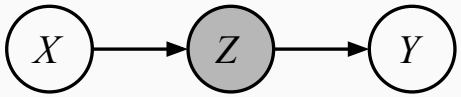


Caminos

- Secuencia de nodos entre un miembro de X y un miembro de Y tal que cada par de nodos adyacente está conectado por una arista sin importar la dirección

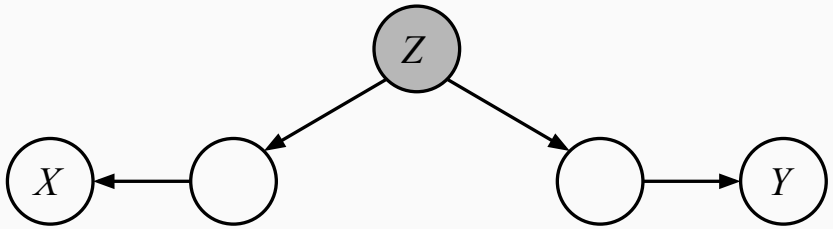
Caminos bloqueados

- Dado un conjunto de nodos Z, se dice que el camino está bloqueado si hay un nodo z para el cual se cumple al menos que
  - z está en Z y tiene una arista sobre el camino entra y la otra sale (cadena)



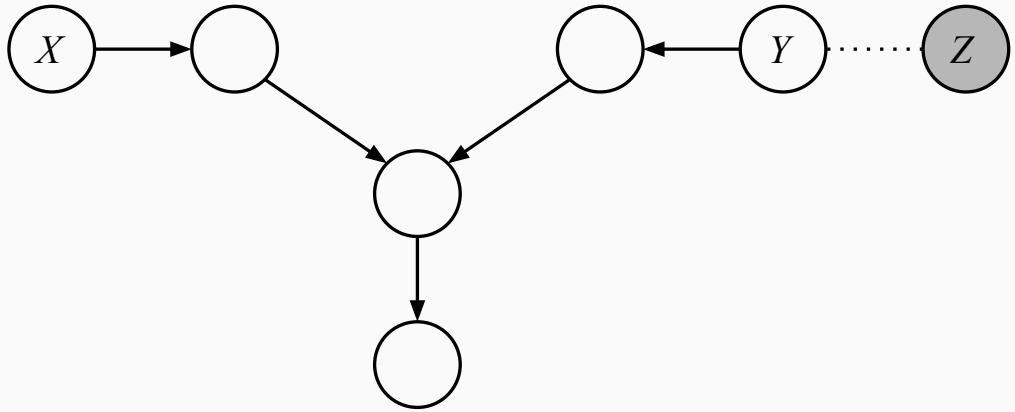
Caminos bloqueados

- Dado un conjunto de nodos Z, se dice que el camino está bloqueado si hay un nodo z para el cual se cumple al menos que
  - z está en Z y tiene una arista sobre el camino entra y la otra sale (cadena)
  - z está en Z y ambas aristas salen (causa común)



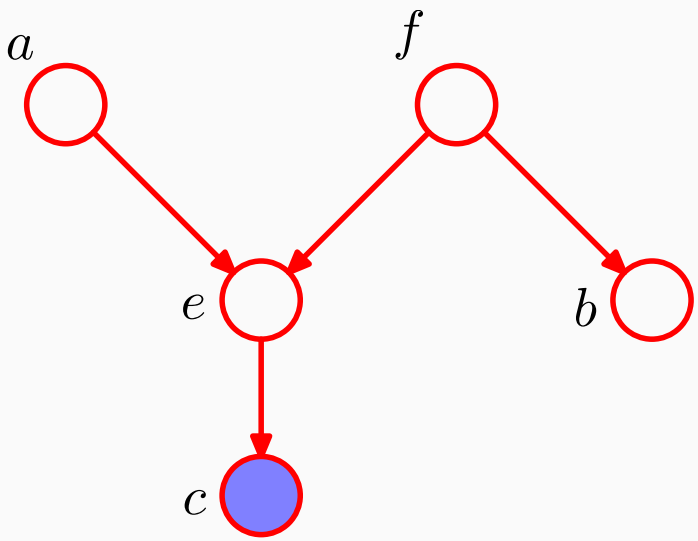
Caminos bloqueados

- Dado un conjunto de nodos Z, se dice que el camino está bloqueado si hay un nodo z para el cual se cumple al menos que
  - z está en Z y tiene una arista sobre el camino entra y la otra sale (cadena)
  - z está en Z y ambas aristas salen (causa común)
  - z ni sus descendientes está en Z y ambos caminos entran a z (efecto común)



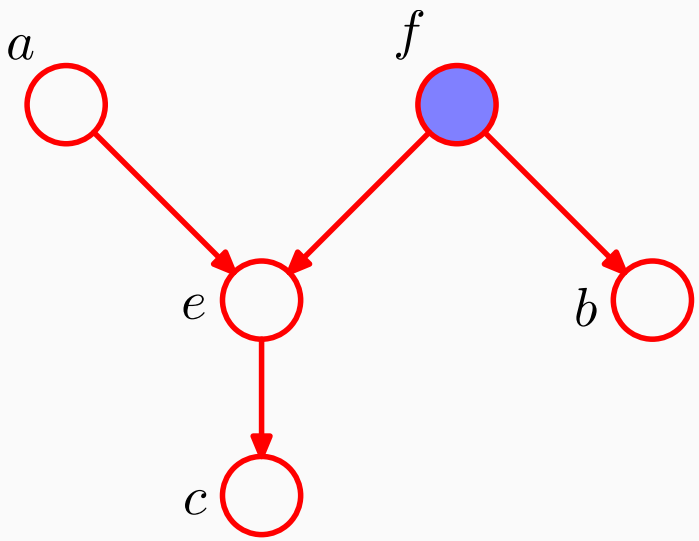
Separación D: ejemplo

$a \not\perp\!\!\!\perp b \mid c$



Tomada de PRML (Bishop 2009)

$a \perp\!\!\!\perp b \mid f$

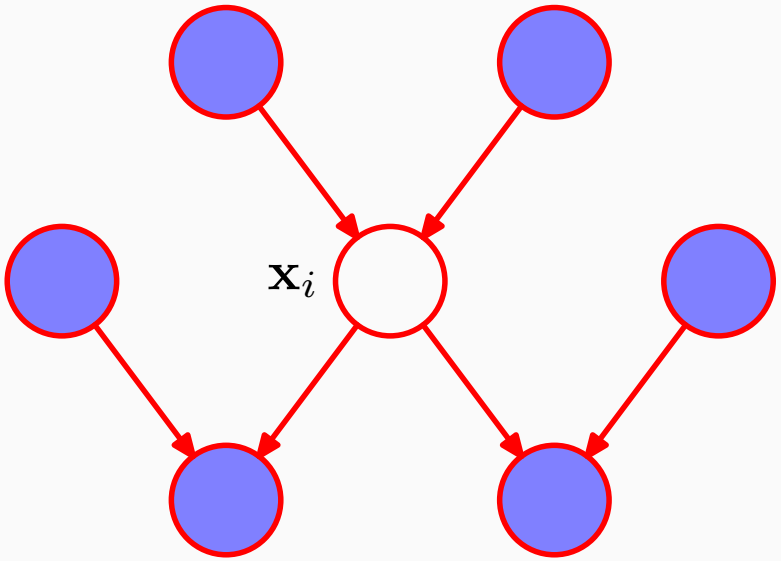


Tomada de PRML (Bishop 2009)

Cobija de Markov

- Conjunto de nodos Y que hacen un nodo x independiente de cualquier otro nodo z en el grafo: padres, hijos y otros padres de hijos

$$P(x|Y,z) = P(x|Y)$$



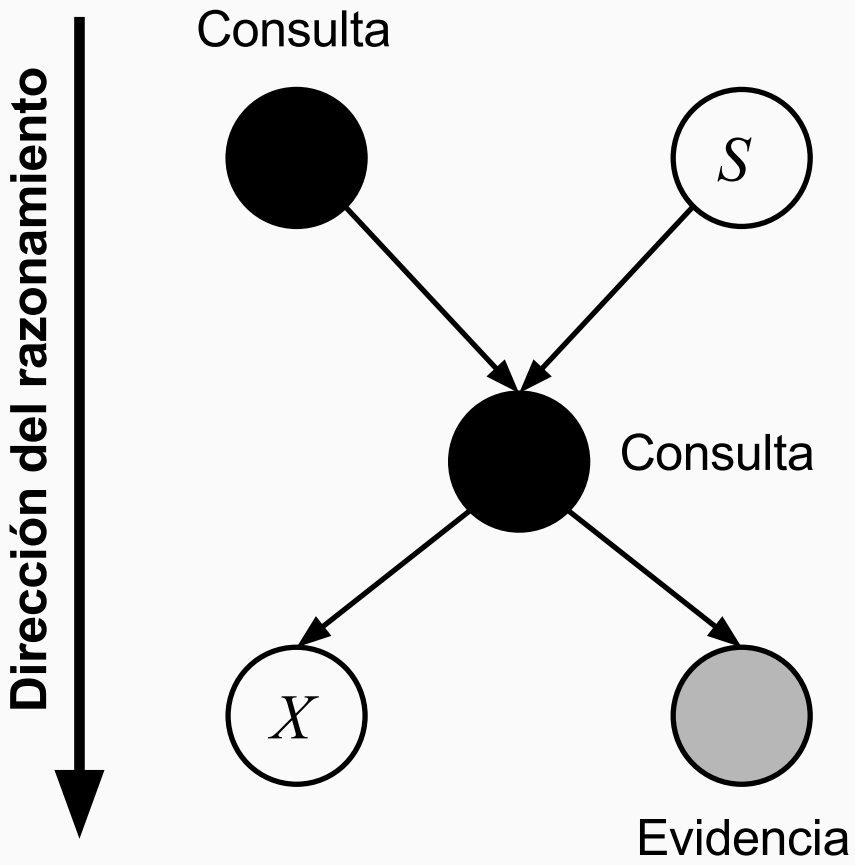
Tomada de PRML (Bishop 2009)

Inferencia en modelos gráficos

- Exacta
  - Intratable para distribuciones y topologías generales
  - Eficiente para algunas topologías (por ej. árboles, poliárboles, etc.)
  - Eliminación de variables, propagación de creencias, etc.
- Aproximada
  - Muestreo directo
  - Muestreo por rechazo
  - Pesado de verosimilitud
  - Montecarlo por cadenas de markov

Tipos de inferencia en redes bayesianas

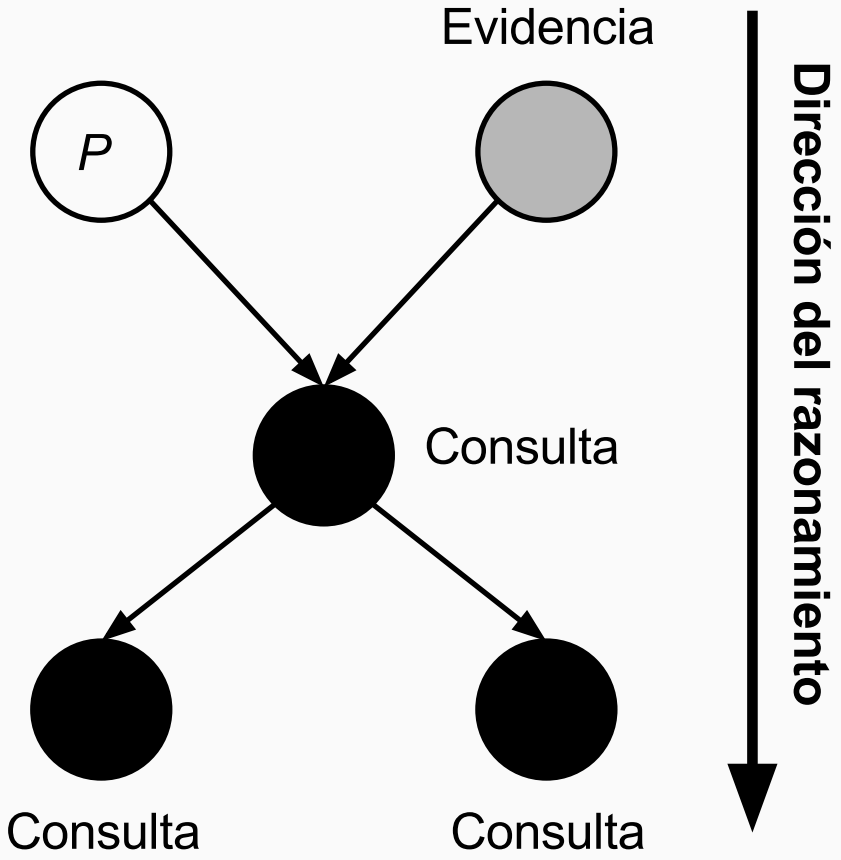
- Diagnóstico



Basada en figura de BAI (Korb y Nicholson 2010)

Tipos de inferencia en redes bayesianas

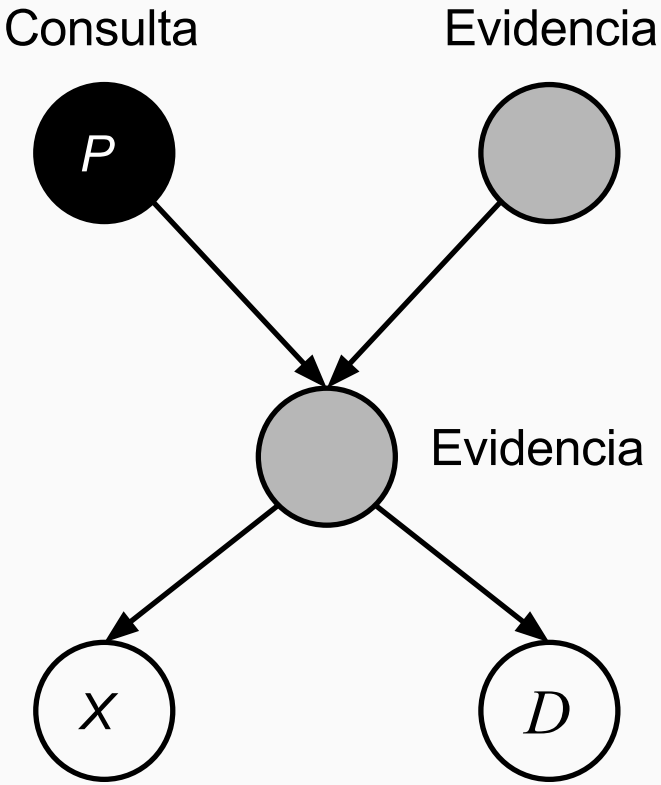
- Predictiva



Basada en figura de BAI (Korb y Nicholson 2010)

Tipos de inferencia en redes bayesianas

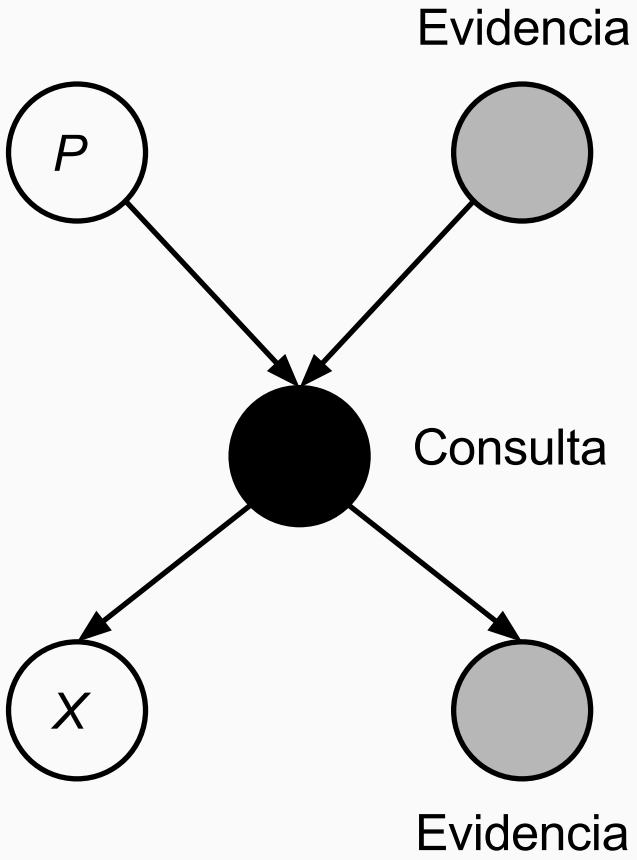
- Intercausal (justificación)



Basada en figura de BAI (Korb y Nicholson 2010)

Tipos de inferencia en redes bayesianas

- Combinada



Basada en figura de BAI (Korb y Nicholson 2010)

Tipos de evidencia

- **Específica**: la variable toma un valor particular (por ej. el paciente es fumador)

Tipos de evidencia

- **Específica**: la variable toma un valor particular (por ej. el paciente es fumador)

- **Negativa**: la variable puede tomar un subconjunto de valores, descartando los demás (por ej. la polución no es alta o la polución es baja o media)

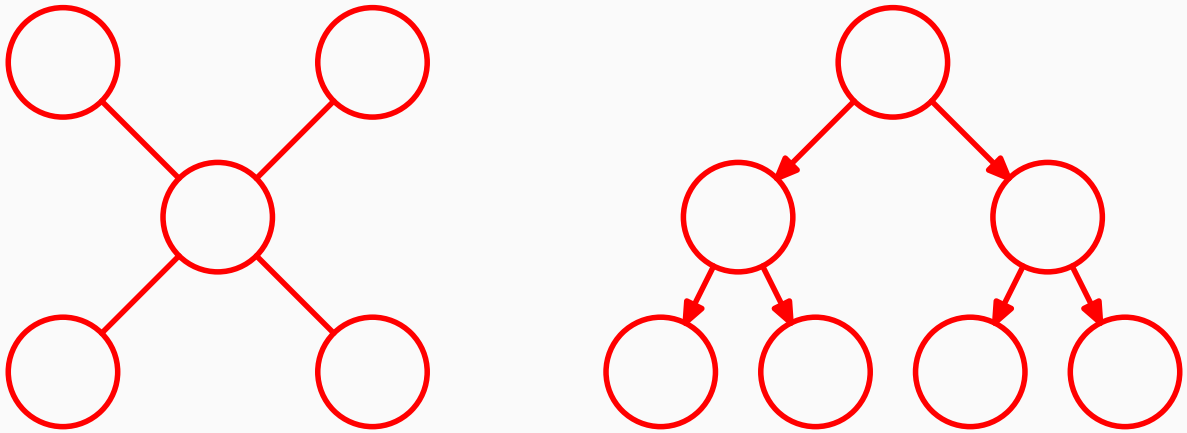
Tipos de evidencia

- **Específica**: la variable toma un valor particular (por ej. el paciente es fumador)
- **Negativa**: la variable puede tomar un subconjunto de valores, descartando los demás (por ej. la polución no es alta o la polución es baja o media)
- **Evidencia virtual**: existe incertidumbre sobre el valor de la variable (por ej. 80 % seguro que los rayos X son positivos)

Árboles

- Grafo donde existe un camino único entre cualquier par de nodos

- Para grafos dirigidos: cada nodo tiene un sólo padre

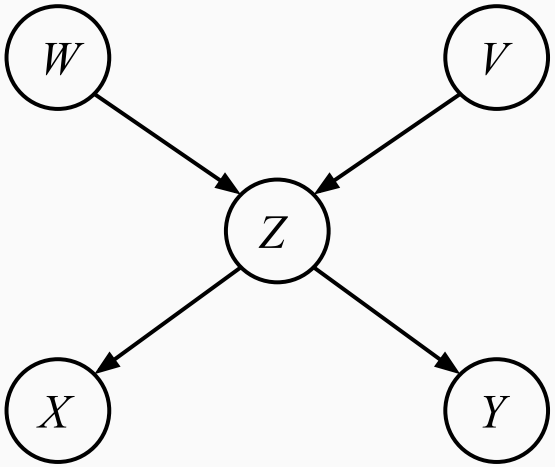


Tomada de PRML (Bishop 2009)

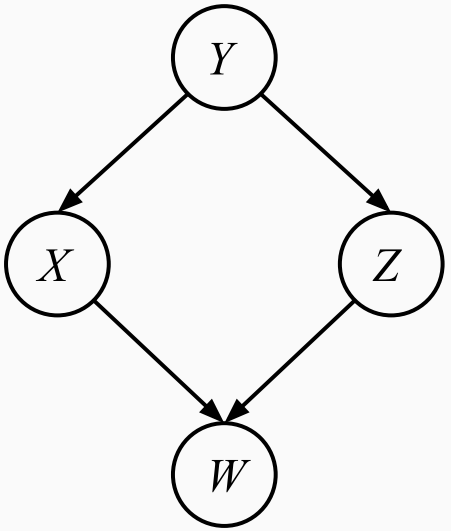
Poliárboles

- Grafos dirigidos con nodos con más de un padre pero tienen un camino único entre cualquier par de nodos

Conexiones únicas



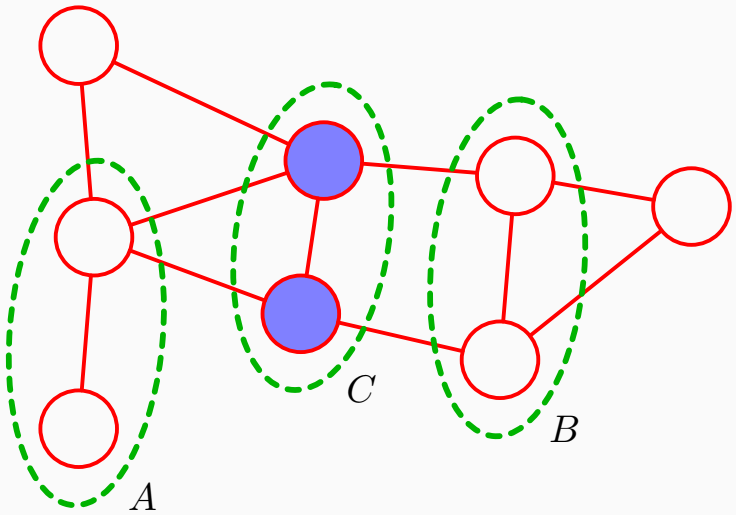
Conexiones múltiples



Redes markovianas o campos aleatorios de Markov (COM)

- Factorización de la distribución conjunta e independencias condicionales se representan con grafos no dirigidos

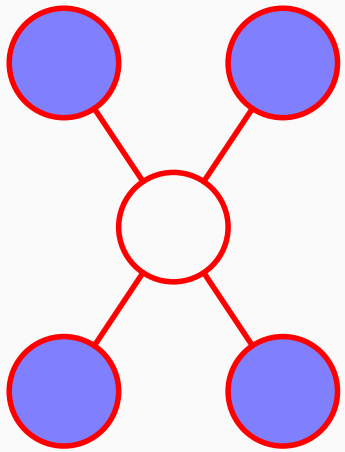
$$A \perp\!\!\!\perp B \mid C$$



Tomada de PRML (Bishop 2009)

Cobija de Markov para redes markovianas

- Cualquier nodo es condicionalmente independiente de cualquier otro nodo en el grafo dado únicamente sus vecinos



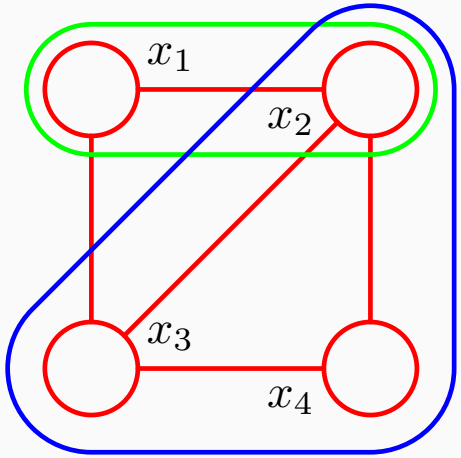
Tomada de PRML (Bishop 2009)

- Distribución conjunta se descompone de acuerdo a funciones sobre los “cliques” del grafo

$$P(x_i, x_j|\mathbf{x}_{\setminus\{i,j\}}) = P(x_i|\mathbf{x}_{\setminus\{i,j\}})P(x_j|\mathbf{x}_{\setminus\{i,j\}})$$

Cliques

- Un **clique** c es un subconjunto de nodos completamente conectados



Tomada de PRML (Bishop 2009)

Probabilidad conjunta en redes markovianas

- $P(\mathbf{x}) > 0$ satisface las propiedades de independencia condicional de un grafo no dirigido $\mathcal{G}$ si y sólo si puede representarse como un producto de factores, uno por clique máximo, es decir

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{\mathcal{C}} \psi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}})$$

donde $\mathcal{C}$ es el conjunto de todos los cliques máximos de $\mathcal{G}$ y Z es la función de partición dada por

$$Z \triangleq \sum_{\mathbf{x}} \prod_{\mathcal{C}} \psi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}})$$

Funciones potencial

- Como $P(\mathbf{x}) > 0$, las funciones potencial se pueden expresar como exponenciales

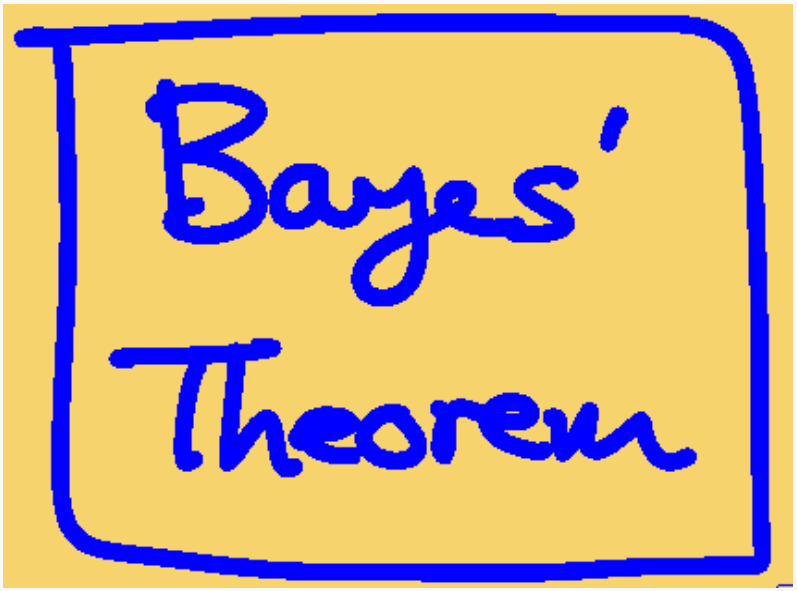
$$\psi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}}) = exp(-E(\mathbf{x}_{\mathcal{C}}))$$

  donde $E(\mathbf{x}_{\mathcal{C}})$ es la función de energía y la representación exponencial se conoce como distribución de Boltzmann

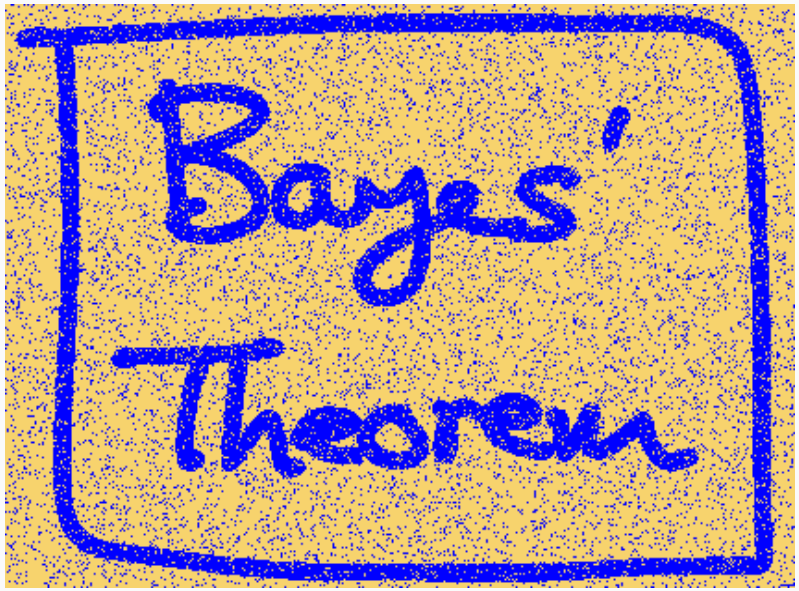
- La distribución conjunta está dada por el producto de funciones potencial, por lo que la energía total se obtiene sumando las energías de cada clique máximo

Red markoviana para quitar ruido en una imagen binaria (1)

Original ($x_i \in \{-1, +1\}$)



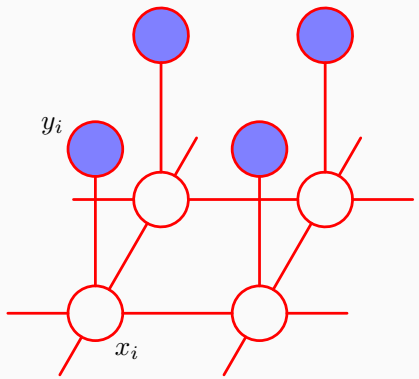
Ruidosa ($y_i \in \{-1, +1\}$)



Imágenes tomadas de Bishop, *Pattern Recognition and Machine Learning*, 2006.

Red markoviana para quitar ruido en una imagen binaria (2)

- Presuposiciones
  - Ruido se genera cambiando el signo de la imagen original
  - Signo de pixel x_i en imagen original se correlaciona con el de sus vecinos x_j y con el del pixel y_i de imagen ruidosa
- Red markoviana
  - Nodo corresponden a pixeles en imagen original y ruidosa
  - Cliques máximos entre cada par de pixeles vecinos ($\{x_i, x_j\}$) y entre cada pixel de imagen original y ruidosa ($\{x_i, y_i\}$)



Imágenes tomadas de Bishop, *Pattern Recognition and Machine Learning*, 2006.

Red markoviana para quitar ruido en una imagen binaria (3)

- Funciones de energía
  - Para $\{x_i, x_j\}$: $-\beta x_i x_j$
  - Para $\{x_i, y_i\}$: $-\eta x_i y_i$
  - Preferencia a un signo: $h x_i$
- Energía total

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

- Distribución conjunta

$$P(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} e^{-E(\mathbf{x},\mathbf{y})}$$

Red markoviana para quitar ruido en una imagen binaria (4)

- Algoritmo *iterated conditional modes* (ICM) para obtener $P(\mathbf{x}|\mathbf{y})$
  1. Inicializa $x_i = y_i$
  2. Calcula la energía dado que un nodo $x_i = +1$ y $x_i = -1$ y fija el valor con menor energía.
  3. Repite 2 hasta cumplir criterio de paro

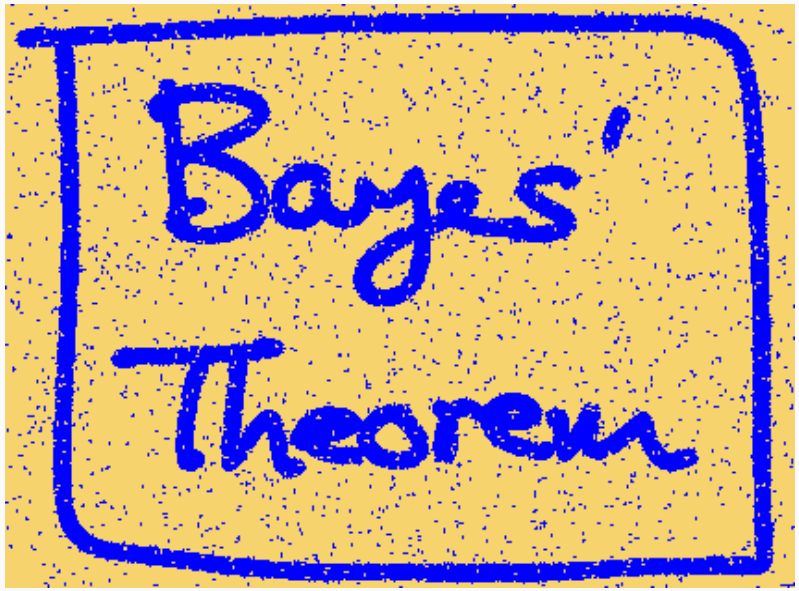


Imagen tomada de Bishop, *Pattern Recognition and Machine Learning*, 2006.

Conversión de redes bayesianas a markovianas

- Nos debemos asegurar que el conjunto de variables que aparecen en cada distribución condicional sean miembros de al menos un clique

- Para nodos de una red bayesiana con más de un padre es necesario agregar aristas entre los nodos padre en la red markoviana (**moralización**)

Conversión de redes bayesianas a markovianas

- Cadenas en redes bayesianas

$$P(\mathbf{x}) = P(x_1)P(x_2|x_1)P(x_3|x_2)\cdots P(x_n|x_{n-1})$$

Tomada de PRML (Bishop 2009)

Conversión de redes bayesianas a markovianas

- Cadenas en redes bayesianas

$$P(\mathbf{x}) = P(x_1)P(x_2|x_1)P(x_3|x_2)\cdots P(x_n|x_{n-1})$$

Tomada de PRML (Bishop 2009)

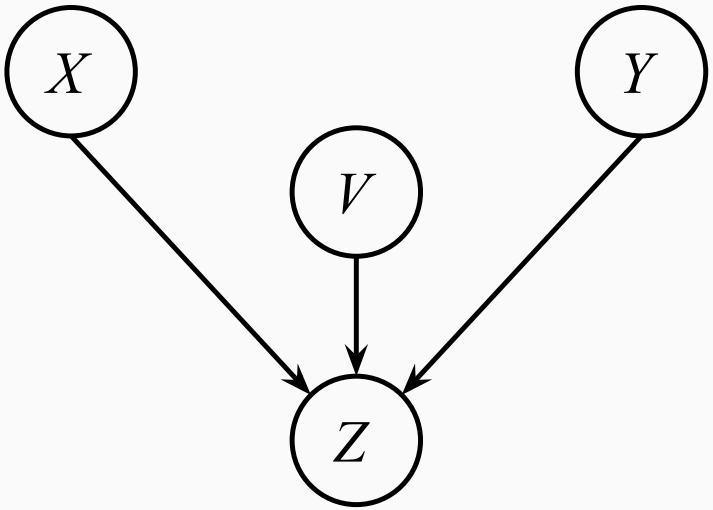
- Cadenas en redes markovianas

$$P(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{n,n-1}(x_{n-1}, X_n)$$

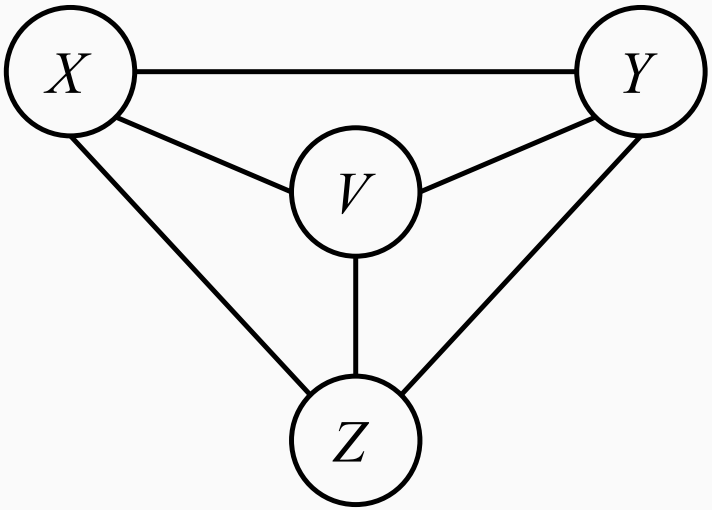
Tomada de PRML (Bishop 2009)

Conversión: otra topología

Red bayesiana

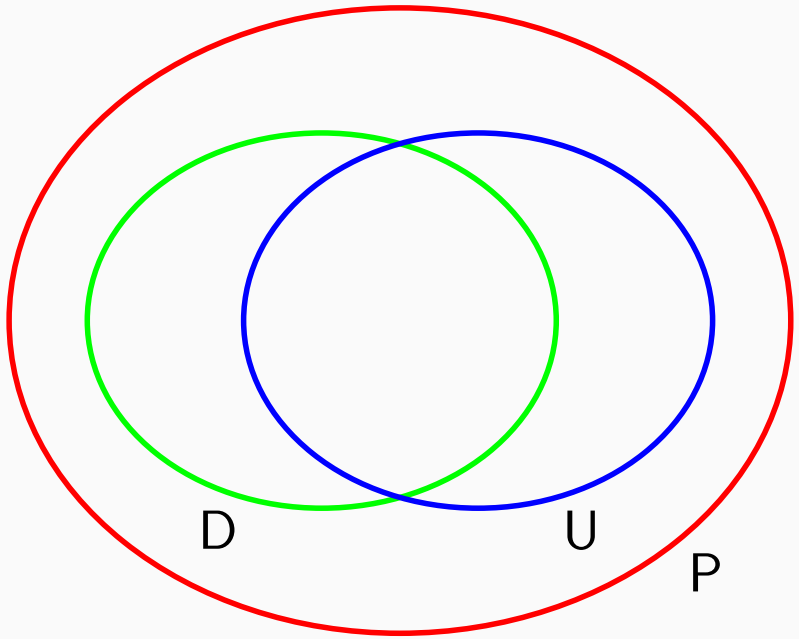


Red markoviana



¿Que distribuciones podemos representar?

- Redes bayesianas vs redes markovianas



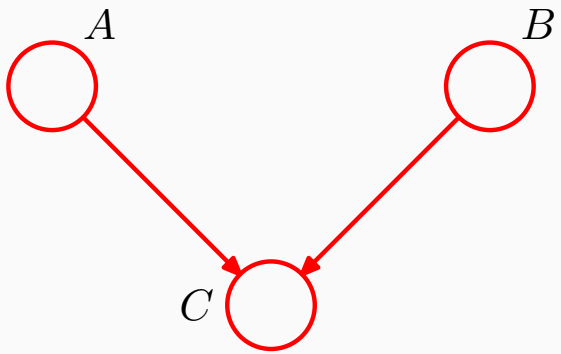
Tomada de PRML (Bishop 2009)

Conversión: limitaciones

- Algunas redes bayesianas no se pueden representar como redes markovianas

$$A \perp\!\!\!\perp B \mid \varnothing$$

$$A \not\perp\!\!\!\perp B \mid C$$



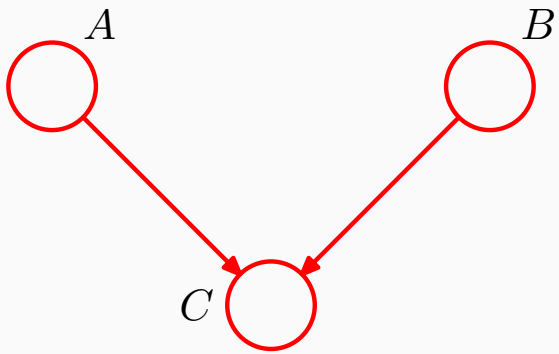
Tomada de PRML (Bishop 2009)

Conversión: limitaciones

- Algunas redes bayesianas no se pueden representar como redes markovianas

$$A \perp\!\!\!\perp B \mid \varnothing$$

$$A \not\perp\!\!\!\perp B \mid C$$



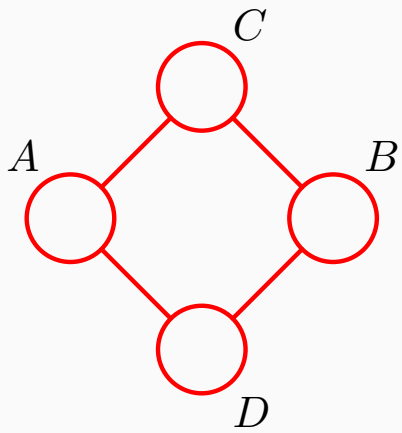
Tomada de PRML (Bishop 2009)

- Y viceversa

$$A \not\perp\!\!\!\perp B \mid \varnothing$$

$$C \perp\!\!\!\perp D \mid A \cup B$$

$$A \perp\!\!\!\perp B \mid C \cup D$$



Tomada de PRML (Bishop 2009)