2016-2017

Aprendizaje Automático

5. Modelos gráficos

etsinf
Escola Tècnica Superior d'Enginyeria Informàtica

Francisco Casacuberta Nolla
(fcn@dsic.upv.es)

Enrique Vidal Ruiz
(evidal@dsic.upv.es)

Departament de Sistemas Informàtics i Computació (DSIC)

Universitat Politècnica de València (UPV)

Septiembre, 2016

Index

Modelos gráficos (MG)

- Los MGs y concretamente las redes bayesianas constituyeron la aproximación probabilística a la IA. Uno de los más famososs impulsores fue Judea Pearl ganador del "ACM A.M. Turing Award" en 2011.
- Concepto: Representación compacta de distribuciones de probabilidad conjunta mediante grados dirigidos (redes bayesianas) y no dirigidos (campos aleatorios markovianos) (teoría de grafos + teoría de la probabilidad). Los MGs generalizan a las redes neuronales y a los modelos de Markov ocultos entre otros.
- Aspectos:
  - Inferencia: para responder cuestiones sobre la distribución de probabilidad
  - Aprendizaje: para obtener los parámetros y la estructura del modelo gráfico
- Aplicaciones:
  - Diagnóstico médico, de fallos, ...
  - Visión por computador: segmentación de imágenes, reconstrucción 3D, análisis de escenas
  - Procesado del lenguaje natural: reconocimiento del habla, extracción de información textual, traducción automática, ...
  - Robótica: planificación, localización, ...

PÁGINA INTENCIONADAMENTE EN BLANCO

Algunos conceptos sobre la teoría de las probabilidades

Probabilidad $P(x): \quad \sum_x P(x) = 1$

Probabilidad conjunta $P(x,y): \quad \sum_x \sum_y P(x,y) = 1$

Probabilidad condicional $P(x \mid y): \quad \sum_x P(x \mid y) = 1 \quad \forall y$

Marginales $P(x) \; = \; \sum_y P(x,y), \qquad P(y) \; = \; \sum_x P(x,y)$

Regla de la probabilidad conjunta $P(x,y) \; = \; P(x)\, P(y \mid x)$

Regla de la cadena $P(x_1, x_2, \ldots, x_N) \; = \; P(x_1) \prod_{i=2}^{N} P(x_i \mid x_1, \ldots, x_{i-1})$

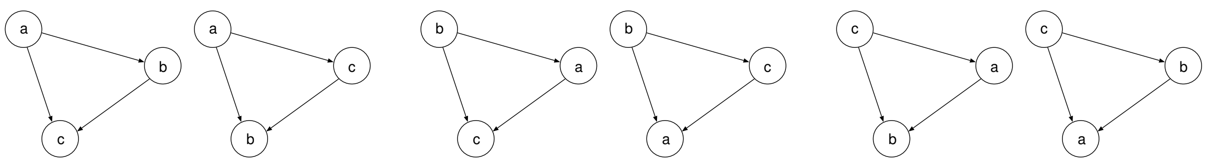
Regla de Bayes $P(y \mid x)\; P(x) \; = \; P(y)\, P(x \mid y)$

Factorización de distribuciones conjuntas

Una distribución conjunta sobre *tres* variables puede expresarse exactamente mediante *seis* factorizaciones completas diferentes:

$$
\begin{aligned}
P(a,b,c) \; &= \; P(a)\; P(b \mid a)\; P(c \mid a,b) \; = \; P(a)\; P(c \mid a)\; P(b \mid a,c) \\
&= \; P(b)\; P(a \mid b)\; P(c \mid a,b) \; = \; P(b)\; P(c \mid b)\; P(a \mid b,c) \\
&= \; P(c)\; P(a \mid c)\; P(b \mid a,c) \; = \; P(c)\; P(b \mid c)\; P(a \mid b,c)
\end{aligned}
$$

Estas factorizaciones se pueden representar mediante grafos dirigidos acíclicos:



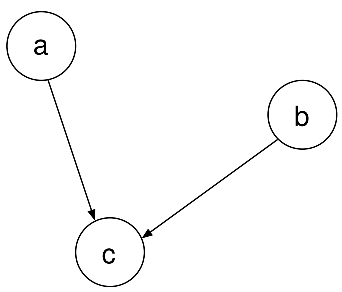
Index

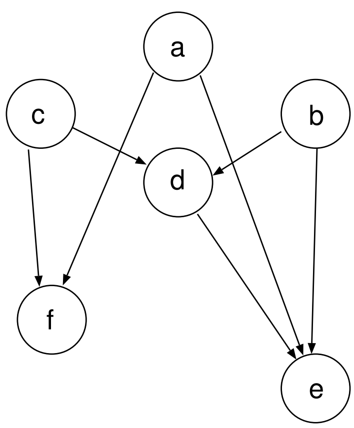
Redes bayesianas: ejemplos

Si hay dependencias inexistentes (o despreciables), la factorización exacta (o aproximada) de una distribución conjunta puede ser incompleta, lo que queda reflejado en el grafo correspondiente. Ejemplos:

$$P(a,b,c) = P(a)\ P(b)\ P(c \mid a,b)$$

$$P(a,b,c,d,e,f) = P(a)\ P(b)\ P(c)\ P(d \mid b,c)\ P(e \mid a,b,d)\ P(f \mid a,c)$$

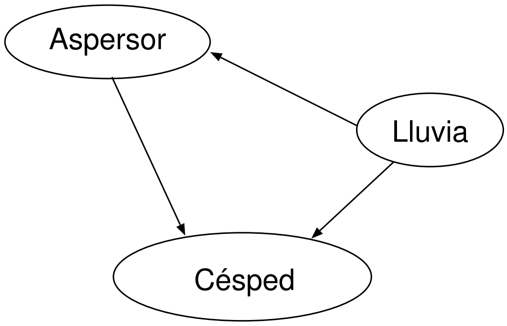




Redes bayesianas: un ejemplo detallado

$P(A \mid L)$

Lluvia (L)	Aspersor (A) f	p
s	0.01	0.99
n	0.40	0.60



$P(L)$

Lluvia (L) s	n
0.2	0.8

Aspersor	f:	funciona
	p:	parado
Césped	m:	mojado
	r:	reseco
Lluvia	s:	sí llueve
	n:	no llueve

$P(C \mid A, L)$

Lluvia (L)	Aspersor (A)	Césped (C) m	r
s	f	0.99	0.01
s	p	0.80	0.20
n	f	0.90	0.10
n	p	0.00	1.00

Distribución conjunta: $P(L, A, C) = P(L)\ P(A \mid L) P(C \mid L, A)$

Ejercicio: calcular $P(L = l, A = a, C = c), \quad l \in \{\mathrm{s}, \mathrm{n}\}, a \in \{\mathrm{f}, \mathrm{p}\}, c \in \{\mathrm{m}, \mathrm{r}\}.$

Un ejemplo detallado (cont.)

- ¿Cuál es la probabilidad de que llueva si el césped está mojado?

$$
\begin{aligned}
P(L = \mathrm{s} \mid C = \mathrm{m}) &= \frac{P(L = \mathrm{s}, C = \mathrm{m})}{P(C = \mathrm{m})} \\
&= \frac{\sum_{a \in \{\mathrm{f},\mathrm{p}\}} P(L = \mathrm{s}, A = a, C = \mathrm{m})}{\sum_{a \in \{\mathrm{f},\mathrm{p}\}, l \in \{\mathrm{s},\mathrm{n}\}} P(L = l, A = a, C = \mathrm{m})} \\
&= \frac{0.00198 + 0.1584}{0.00198 + 0.288 + 0.1584 + 0.0} \\
&= 0.3577
\end{aligned}
$$

- El césped está mojado, ¿llueve?

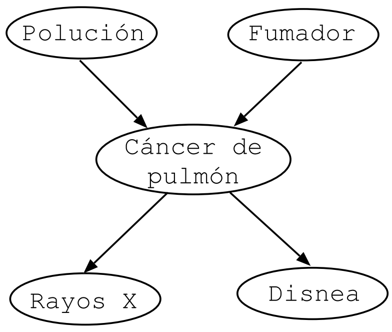
$$\arg\max_{l \in \{\mathrm{s},\mathrm{n}\}} P(L = l \mid C = \mathrm{m}) \ = \ \mathrm{n}$$

Ejercicio: a) Calcular $P(A = a \mid L = l, C = c), \quad a \in \{\mathrm{p}, \mathrm{f}\}, l \in \{\mathrm{s}, \mathrm{n}\}, c \in \{\mathrm{m}, \mathrm{r}\}.$
b) Llueve y el cesped está mojado, ¿cómo está el aspersor?

Redes bayesianas: otro ejemplo

$P(P)$

Polución (P)	
a	b
0.1	0.9



$P(F)$

Fumador (F)	
s	n
0.3	0.7

$P(X \mid C)$

Cáncer (C)	Rayos X (X) p	Rayos X (X) n
p	0.90	0.10
n	0.20	0.80

$P(D \mid C)$

Cáncer (C)	Disnea (D) s	Disnea (D) n
p	0.65	0.35
n	0.30	0.70

Polución	a:	alto
	b:	bajo
Fumador	s:	sí
	n:	no
Disnea	s:	sí
	n:	no
Rayos X	p:	positivo
	n:	negativo
Cáncer	p:	positivo
	n:	negativo

$P(C \mid P, F)$

Polución (P)	Fumador (F)	Cáncer (C) p	Cáncer (C) n
a	s	0.08	0.92
a	n	0.05	0.95
b	s	0.03	0.97
b	n	0.001	0.999

Ejercicio: ¿Cuál es la probabilidad de que un paciente no fumador no tenga cáncer si la radiografía ha dado un resultado negativo pero sufre de disnea?

Redes bayesianas

Una red bayesiana es un grafo dirigido y acíclico (“directed acyclic graph” -DAG-) donde:

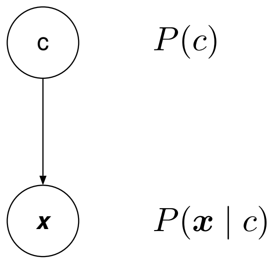
- Los nodos representan:
  - variables (discretas o contínuas)
  - distribuciones de probabilidad condicional para cada variable x_i dados los valores de las variables asociadas a los nodos padres $a(x_i)$
- Los arcos representan dependencias entre las variables

Una red bayesiana con nodos $x_1, \ldots, x_D$ define una distribución de probabilidad conjunta:

$$P(x_1, \ldots, x_D) = \prod_{i=1}^{D} P(x_i \mid a(x_i))$$

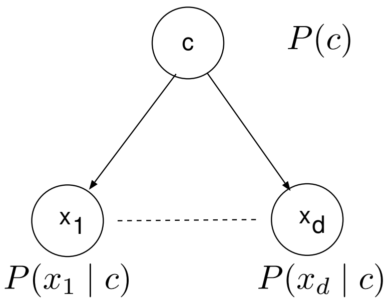
Algunas redes bayesianas simples

- El *clasificador de Bayes* ($\boldsymbol{x} \in \mathbb{R}^d$ y $c \in \{1, \ldots, C\}$):



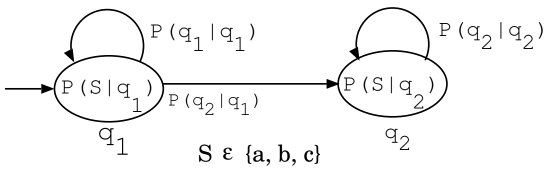
$$P(\boldsymbol{x}, c) = P(c)\, P(\boldsymbol{x} \mid c)$$
$$P(c \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{\sum_c P(\boldsymbol{x}, c)}$$

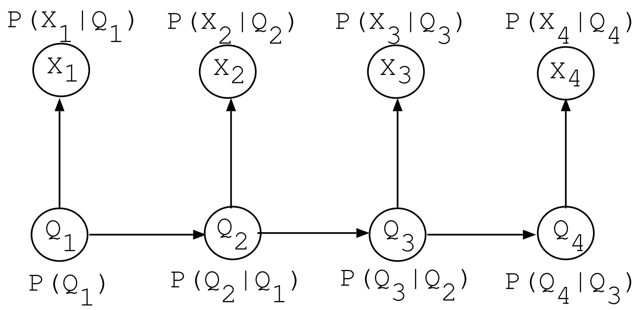
- *Modelos naive-Bayes* ($x_i \in \mathbb{R}$ con $1 \leq i \leq d$ y $c \in \{1, \ldots, C\}$):



$$P(x_1, \ldots, x_d, c) = P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

Otro ejemplo de red bayesiana: modelo oculto de Markov





$$P(X_1\, X_2\, X_3\, X_4, Q_1\, Q_2\, Q_3\, Q_4) =$$

$$P(Q_1)\, P(X_1 \mid Q_1)\, P(Q_2 \mid Q_1)\, P(X_2 \mid Q_2)\, P(Q_3 \mid Q_2)\, P(X_3 \mid Q_3)\, P(Q_4 \mid Q_3)\, P(X_4 \mid Q_4)$$

$$P(X_1 = a, X_2 = a, X_3 = b, X_4 = c) =$$

$$\sum_{r_1, r_2, r_3, r_4 \in \{q_1, q_2\}} P(X_1 = a, X_2 = a, X_3 = b, X_4 = c, Q_1 = r_1, Q_2 = r_2, Q_3 = r_3, Q_4 = r_4)$$

Index

Independencia condicional

Se dice que a es condicionalmente independiente de b dado c (o también que a está D-separado de b por c, y se denota como: $a \perp\!\!\!\perp b \mid c$) si:

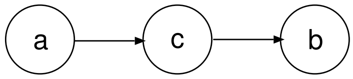
$$P(a \mid b, c) = P(a \mid c) \quad \Leftrightarrow \quad P(a, b \mid c) = P(a \mid c) P(b \mid c)$$

Se dice que a es incondicionalmente independiente de b (y se denota como: $a \perp\!\!\!\perp b \mid \emptyset$) si:

$$P(a, b) = P(a)\, P(b)$$

Independencia in/condicional: estructuras típicas de RBs (I)

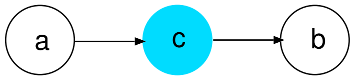
- Caso 1: ¿Cuál es la relación entre a y b?



$$P(a,b,c) = P(a)P(c \mid a)P(b \mid c)$$

$$P(a,b) = \sum_c P(a)P(c \mid a)P(b \mid c)$$

En general $P(a,b) \neq P(a)P(b) \Rightarrow a \not\perp\!\!\!\perp b \mid \emptyset$

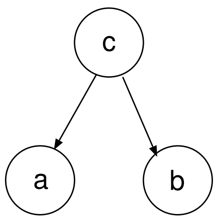


$$P(a,b \mid c) = \frac{P(a,b,c)}{P(c)} = \frac{P(a)P(c \mid a)P(b \mid c)}{P(c)}$$
$$= P(a \mid c)P(b \mid c) \;\Rightarrow\; a \perp\!\!\!\perp b \mid c$$

El nodo c es cabeza-con-cola con respecto al camino desde a a b via c: Dos nodos a y b conectados via un nodo c cabeza-con-cola son independientes condicionalmente si c está dado (nodo c bloquea al camino desde a a b).

Independencia in/condicional: estructuras típicas de RBs (II)

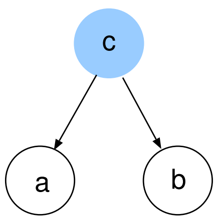
- Caso 2: ¿Cuál es la relación entre a y b?



$$P(a,b,c) = P(a \mid c)P(b \mid c)P(c)$$

$$P(a,b) = \sum_c P(a \mid c)P(b \mid c)P(c)$$

En general $P(a,b) \neq P(a)P(b) \Rightarrow a \not\perp\!\!\!\perp b \mid \emptyset$

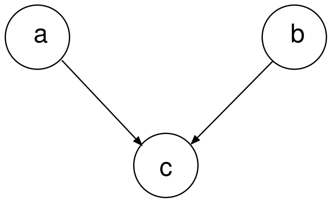


$$P(a,b \mid c) = \frac{P(a,b,c)}{P(c)} = P(a \mid c)P(b \mid c) \Rightarrow a \perp\!\!\!\perp b \mid c$$

El nodo c es cola-con-cola con respecto a el camino desde a a b via c: Dos nodos a y b conectados via un nodo c cola-con-cola son independientes condicionalmente si c está dado, (nodo c *bloquea* al camino desde a a b).

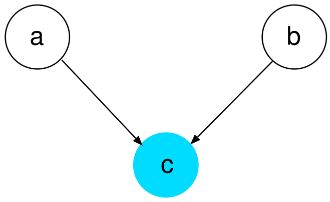
Independencia in/condicional: estructuras típicas de RBs (III)

- Caso 3: ¿Cuál es la relación entre a y b?



$$P(a,b,c) = P(a)P(b)P(c \mid a,b)$$

$$P(a,b) = \sum_c P(a)P(b)P(c \mid a,b) = P(a)P(b) \Rightarrow a \perp\!\!\!\perp b \mid \emptyset$$

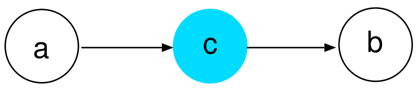


$$P(a,b \mid c) = \frac{P(a,b,c)}{P(c)} = \frac{P(a)P(b)P(c \mid a,b)}{P(c)}$$

En general $P(a,b \mid c) \neq P(a \mid c)P(b \mid c) \Rightarrow a \not\perp\!\!\!\perp b \mid c$

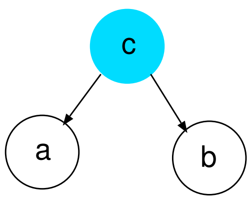
El nodo c es to cabeza-con cabeza con respecto al camino desde a a b via c: Dos nodos a y b conectados via un nodo c cabeza-con-cabeza no son independientes condicionalmete si c está dado, pero en caso contrario son independientes incondicionalmente (nodo c bloquea al camino desde a a b cuando c está dado).

Independencia conditional: Resumen



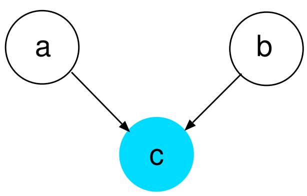
Dirección causal $a \perp\!\!\!\perp b \mid c$

$$P(a,b \mid c) = P(a \mid c)P(b \mid c)$$



Causa común $a \perp\!\!\!\perp b \mid c$

$$P(a,b \mid c) = P(a \mid c)P(b \mid c)$$



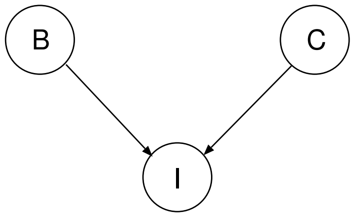
Estructura en V $a \not\perp\!\!\!\perp b \mid c$

En general $P(a,b \mid c) \neq P(a)P(b)$

Independencia condicional: Ejemplo (I)

- B es el estado de la batería (cargada $B = 1$ o descargada $B = 0$)
- C es el estado del depósito de combustible (lleno $C = 1$ o vacío $C = 0$)
- I es el estado del indicador eléctrico del combustible (lleno $I = 1$ o vacío $I = 0$)

$$\begin{aligned} P(B = 1) = P(C = 1) &= 0.9 \\ P(I = 1 \mid B = 1, C = 1) &= 0.8 \\ P(I = 1 \mid B = 1, C = 0) &= 0.2 \\ P(I = 1 \mid B = 0, C = 1) &= 0.2 \\ P(I = 1 \mid B = 0, C = 0) &= 0.1 \end{aligned}$$



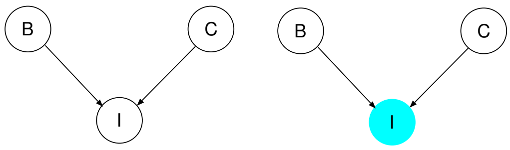
$$P(B, C, I) = P(B)\; P(C)\; P(I \mid B, C)$$

Independencia condicional: Ejemplo (II)

B	C	I = 1	I = 0
1	1	0.8	0.2
1	0	0.2	0.8
0	1	0.2	0.8
0	0	0.1	0.9

B = 1	B = 0
0.9	0.1

C = 1	C = 0
0.9	0.1



$$P(B, C, I) = P(B)\; P(C)\; P(I \mid B, C)$$

Si no tenemos información sobre I ¿cuál es la probabilidad conjunta de B y C?:

$$\begin{aligned} P(B, C) &= P(B, C, I\!=\!0) + P(B, C, I\!=\!1) \\ &= P(B)\; P(C)\; \Big(P(I\!=\!0 \mid B, C) + P(I\!=\!1 \mid B, C)\Big) = P(B)\; P(C) \;\Rightarrow\; B \perp\!\!\!\perp C \mid \emptyset \end{aligned}$$

Supongamos que vemos $I = 0$, ¿Cuál es la probabilidad $P(B = 0, C = 0 \mid I = 0)$?

$$P(B\!=\!0, C\!=\!0 \mid I\!=\!0) = \frac{P(B\!=\!0)\; P(C\!=\!0)\; P(I\!=\!0 \mid B\!=\!0, C\!=\!0)}{\sum_{b,c \in \{0,1\}} P(B = b)\; P(C = c)\; P(I\!=\!0 \mid B = b, C = c)} = 0.02857$$

$$P(B\!=\!0 \mid I\!=\!0) = \frac{\sum_{c \in \{0,1\}} P(B\!=\!0)\; P(C = c)\; P(I\!=\!0 \mid B\!=\!0, C = c)}{\sum_{b,c \in \{0,1\}} P(B = b)\; P(C = c)\; P(I\!=\!0 \mid B = b, C = c))} = 0.25714$$

$$P(C\!=\!0 \mid I\!=\!0) = \frac{\sum_{b \in \{0,1\}} P(B\!=\!b)\; P(C\!=\!0)\; P(I\!=\!0 \mid B\!=\!b, C\!=\!0)}{\sum_{b,c \in \{0,1\}} P(B\!=\!b)\; P(C\!=\!c)\; P(I\!=\!0 \mid B\!=\!b, C\!=\!c))} = 0.25714$$

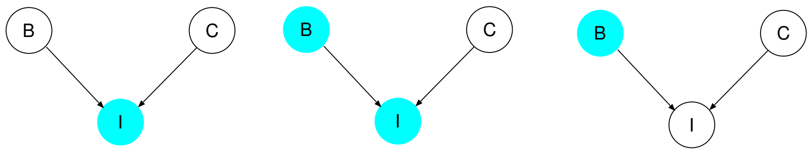
$$P(B\!=\!0, C\!=\!0 \,|\, I\!=\!0) = 0.02857 \neq P(B\!=\!0 \,|\, I\!=\!0)\; P(C\!=\!0 \,|\, I\!=\!0) = 0.06612 \;\Rightarrow\; B \not\perp\!\!\!\perp C \mid I$$

Independencia condicional: Ejemplo (III)

		I	
B	C	1	0
1	1	0.8	0.2
1	0	0.2	0.8
0	1	0.2	0.8
0	0	0.1	0.9

B	
1	0
0.9	0.1

C	
1	0
0.9	0.1



$$P(B, C, I) = P(B)\, P(C)\, P(I \mid B, C)$$

Supongamos que vemos $I = 0$, ¿Cuál es la probabilidad $P(C = 0 \mid I = 0)$?

$$\begin{aligned}
P(C = 0 \mid I = 0) &= \frac{P(I = 0, C = 0)}{P(I = 0)} = \frac{\sum_{B\in\{0,1\}} P(I = 0, C = 0, B)}{\sum_{B\in\{0,1\}} \sum_{C\in\{0,1\}} P(I = 0, C, B)} \\
&= \frac{\sum_{B\in\{0,1\}} P(I = 0 \mid B, C = 0) P(B) P(C = 0)}{\sum_{B\in\{0,1\}} \sum_{C\in\{0,1\}} P(I = 0 \mid B, C) P(B) P(C)} \\
&= \frac{0.9 \cdot 0.1 \cdot 0.1 + 0.8 \cdot 0.9 \cdot 0.1}{0.315} = 0.25714 > P(C = 0) = 0.1
\end{aligned}$$

Si ahora vemos $B = 0$: $P(C = 0 \mid I = 0, B = 0) \approx 0.111 < P(C = 0 \mid I = 0)$,
es decir, C depende de B si se conoce I

Por otra parte, si no conocemos I: $P(C = 0 \mid B = 0) = 0.1 = P(C = 0)$,
es decir, C no depende de B si no se conoce I

Index

Inferencia con redes bayesianas

- En general, el problema consiste en calcular la probabilidad a posteriori de alguna variable x a partir de las distribuciones conjuntas asociadas a una RB, dada alguna evidencia e (como valores dados de alguna otra variable) y sin importar los valores de resto de las variables f:

$$P(x \mid e) = \frac{P(x,e)}{P(e)} \text{ con } P(e) = \sum_{x,f} P(x,e,f) \text{ y } P(x,e) = \sum_{f} P(x,e,f)$$

- El objetivo es calcular eficientemente $P(e)$ y $P(x,e)$
- Por ejemplo: Sea $P(x_1,x_2,x_3,x_4)$ una distribución conjunta dada por

$$P(x_1,x_2,x_3,x_4) = P(x_2)P(x_1 \mid x_2)P(x_3 \mid x_2)P(x_4 \mid x_3)$$

  y se pretende calcular $P(x_3)$. Si $x_i \in X$ para $i = 1,2,3$ verifica que $|X| = n$

  - $P(x_3) = \sum_{x_1,x_2,x_4} P(x_2)P(x_1 \mid x_2)P(x_3 \mid x_2)P(x_4 \mid x_3) \quad \Rightarrow O(n^3)$ operaciones.
  - $P(x_3) = \sum_{x_2} P(x_2)P(x_3 \mid x_2) \sum_{x_1} P(x_1 \mid x_2) \sum_{x_4} P(x_4 \mid x_3)$
    $= \sum_{x_2} P(x_2)P(x_3 \mid x_2) \quad \Rightarrow O(n)$ operaciones.

Inferencia con redes bayesianas

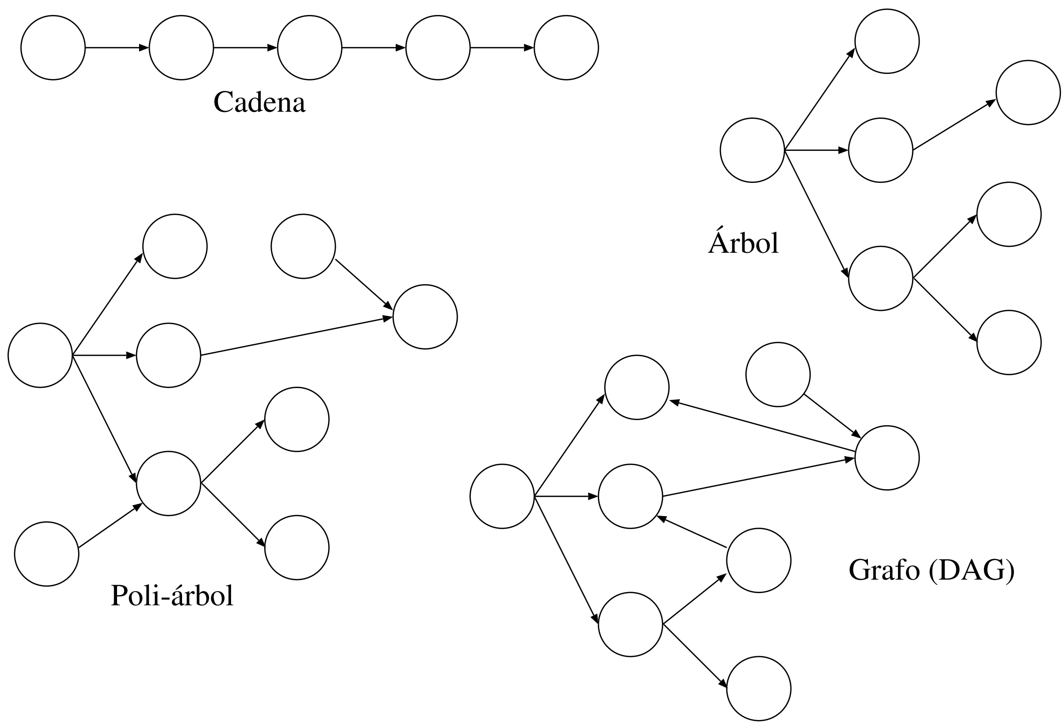
Situaciones donde es útil calcular las probabilidades a-posteriori:

- Predicción: ¿Cuál es la probabilidad de observar un síntoma sabiendo que se tiene una determinada enfermedad?

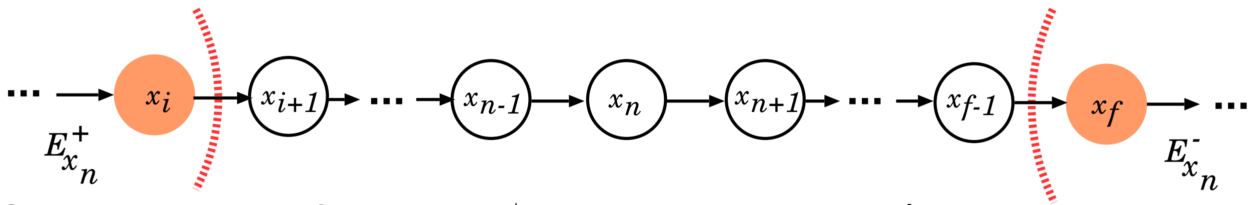
- Diagnóstico: ¿Cuál es la probabilidad de que una determinada enfermedad sea un diagnóstico correcto dados algunos síntomas?

En RB, la dirección de los enlaces entre variables no restringe el tipo de preguntas que se pueden hacer.

Tipos de redes bayesianas



Inferencia en una cadena. Propagación de creencias ("Belief propagation")

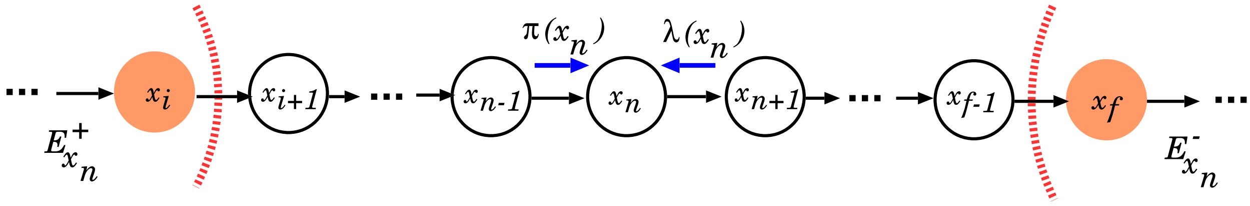


Supongamos que el último $x_i \in E^+_{x_n}$ y el primer $x_f \in E^-_{x_n}$ están dados:

$$
\begin{aligned}
P(x_n \mid x_i, x_f) &= \frac{P(x_i, x_f \mid x_n)\ P(x_n)}{P(x_i, x_f)} \\
&= \frac{P(x_i \mid x_n)\ P(x_f \mid x_n)\ P(x_n)}{P(x_i, x_f)} \qquad \text{(Indepencia condicional)} \\
&= \frac{P(x_n \mid x_i)\ P(x_i)\ P(x_f \mid x_n)\ P(x_n)}{P(x_n)\ P(x_i, x_f)} \qquad \text{(Regla de Bayes)} \\
&= \alpha\ P(x_n \mid x_i)\ P(x_f \mid x_n) \qquad (\alpha = P(x_i)/P(x_i, x_f))
\end{aligned}
$$

- Ejercicio: ¿Qué ocurre si también conocemos $x_{i'} \in E^+_{x_n}$ con $i' < i$?
- Ejercicio: ¿Qué ocurre si también conocemos $x_{f'} \in E^-_{x_n}$ con $f' > f$?

Inferencia en una cadena



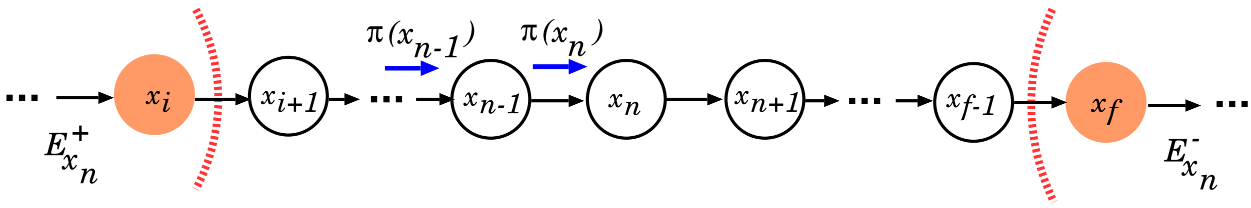
$$P(x_n \mid x_i, x_f) \;=\; \alpha\, P(x_n \mid x_i)\, P(x_f \mid x_n) \;\stackrel{\text{def}}{=}\; \alpha\, \pi(x_n)\, \lambda(x_n)$$

donde $\pi(x_n)$ y $\lambda(x_n)$ se calculan como:

$$\left|\begin{array}{rcl} \pi(x_i) & = & 1 \\ \pi(x_n) & = & \displaystyle\sum_{x_{n-1}} P(x_n \mid x_{n-1})\, \pi(x_{n-1}) \end{array}\right| \begin{array}{rcl} \lambda(x_f) & = & 1 \\ \lambda(x_n) & = & \displaystyle\sum_{x_{n+1}} P(x_{n+1} \mid x_n)\, \lambda(x_{n+1}) \end{array}\right|$$

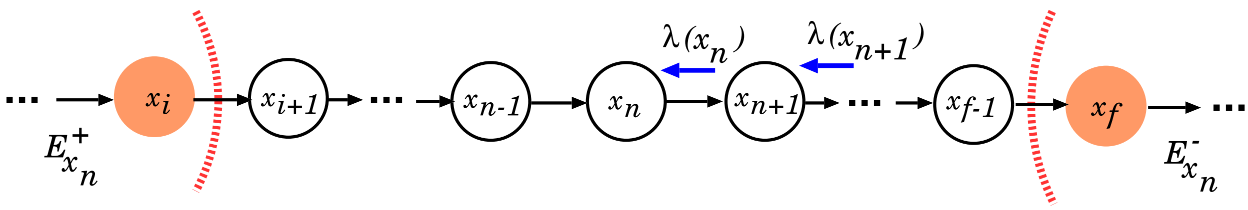
- Ejercicio: $P(x_n)$

Inferencia en una cadena (derivación I)



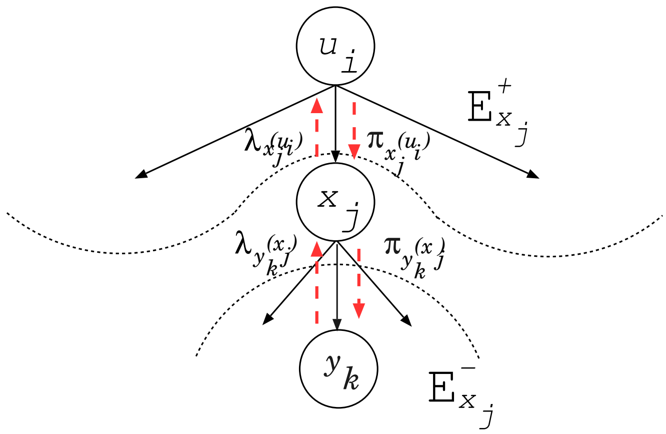
$$\begin{aligned} \pi(x_n) &= P(x_n \mid x_i) = \sum_{x_{n-1}} P(x_n, x_{n-1} \mid x_i) \\ &= \sum_{x_{n-1}} P(x_{n-1} \mid x_i)\, P(x_n \mid x_i, x_{n-1}) = \sum_{x_{n-1}} P(x_{n-1} \mid x_i)\, P(x_n \mid x_{n-1}) \\ &= \sum_{x_{n-1}} \pi(x_{n-1})\, P(x_n \mid x_{n-1}) \\ \pi(x_i) &= 1 \end{aligned}$$

Inferencia en una cadena (derivación II)



$$
\begin{aligned}
\lambda(x_n) &= P(x_f \mid x_n) = \sum_{x_{n+1}} P(x_f, x_{n+1} \mid x_n) \\
&= \sum_{x_{n+1}} P(x_{n+1} \mid x_n)\ P(x_f \mid x_n, x_{n+1}) = \sum_{x_{n+1}} P(x_{n+1} \mid x_n)\ P(x_f \mid x_{n+1}) \\
&= \sum_{x_{n+1}} P(x_{n+1} \mid x_n)\ \lambda(x_{n+1}) \\
\lambda(x_f) &= 1
\end{aligned}
$$

Inferencia en un árbol



Para calcular $P(x_j \mid E^+_{x_j}, E^-_{x_j}) = \alpha\ \pi(x_j)\ \lambda(x_j)$

$$\lambda(x_j) = \prod_{k=1}^{m} \lambda_{y_k}(x_j) \ \text{ con } \ \lambda_{y_k}(x_j) = \sum_{y_k} \lambda(y_k)\ P(y_k \mid x_j)$$

$$\pi(x_j) = \sum_{u_i} P(x_j \mid u_i)\ \pi_{x_j}(u_i) \ \text{ con } \ \pi_{x_j}(u_i) = \alpha \prod_{j' \neq j} \lambda_{x_{j'}}(u_i)\ \pi(u_i)$$

Inferencia en otros tipos de grafos

- Poli-arboles ("Polytrees"). Los nodo pueden tener múltiples padres, pero solo puede existir un camino único entre cualquier par de nodos: una generalización del algoritmo sobre un árbol.
- Grafos generales. Inferencia aproximada:
  - Métodos variacionales.
  - Métodos basados en el muestreo.

Index

Campos de Markov aleatorios

- Aspectos generales:
  - Independencia conditional simplificada
  - Asignación compleja de distribuciones de probabilidad.
- Definición de un campo de Markov aleatorio
  - Dado un conjunto de variables $V = \{x_1, \ldots, x_D\}$
  - Un grafo no-dirigido $R = (V, E)$
  - Para un clique C en R (un subgrafo completamente conectado), V_C es el conjunto variables en el clique C:

$$P(x_1, \ldots, x_D) = \frac{1}{Z} \prod_{\forall C \in R} \psi_C(V_C)$$

$\psi_C(V_C)$ es una función potential (estrictamente positiva) y Z es un factor de normalización (función de partición).

Campos de Markov aleatorios

$$P(x_1, \ldots, x_D) = \frac{1}{Z} \prod_{\forall C \in R} \psi_C(V_C)$$

- Las funciones de potencial son de la forma:

$$\psi_C(V_C) = \exp\left(-E(V_C)\right),$$

donde $E(V_C)$ es una función de energía. Por lo tanto,

$$P(x_1, \ldots, x_D) = \frac{1}{Z} \exp\left(-\sum_{\forall C \in R} E(V_C)\right)$$

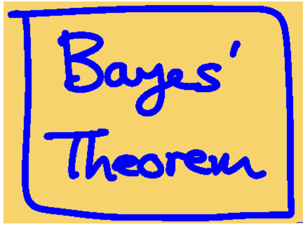
- Un tipo de función de energía interesante es el queda definido mediante funciones lineales generalizadas:

$$E(V_C) = -\sum_k \theta_{C,k}\, f_{C,k}(V_C)$$

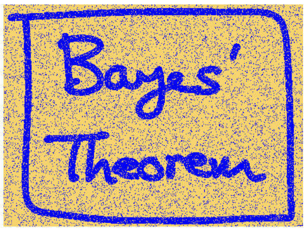
- A partir de una red bayesiana se puede construir un campo de Markov aleatorio

Un ejemplo

Un conjunto $\boldsymbol{x}$ de píxeles binarios
$x_i \in \{-1, +1\}, \;\; 1 \leq i \leq D$



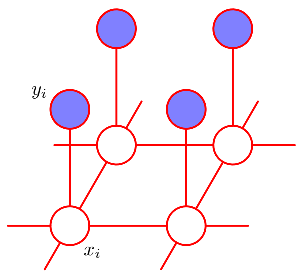
La imagen está corrupta con probabilidad 10% en un conjunto $\boldsymbol{y}$ de píxeles binarios
$y_i \in \{-1, +1\}, \;\; 1 \leq i \leq D$



El objetivo es recuperar la imagen original a partir de la imagen corrupta

Un ejemplo

(Bishop. Pattern Recognition and Machine Learning. 2006)



- Fuerte correlación entre x_i y y_i
- Correlación entre x_i y x_j si los dos píxeles son vecinos.
- La función de energía (cliques máximos: $C_{ij} = (x_i, x_{j \in N(i)})$, $C_i = (x_i, y_i)$ $\forall i, j$):

$$\left. \begin{array}{r} E(V_{C_{ij}}) = -\beta \; x_i \; x_j \\ E(V_{C_i}) = -\nu \; x_i \; y_j \end{array} \right\} \;\; \rightarrow \;\; \sum_{\forall C \in R} E(V_C) = -\beta \sum_{i,j} x_i \; x_j - \nu \sum_i x_i \; y_i$$

- La distribución conjunta:

$$P(x_1, \ldots, x_D, y_1, \ldots, y_D) = \frac{1}{Z} \exp\left(\beta \sum_{i,j} x_i \; x_j + \nu \sum_i x_i \; y_i \right)$$

Un ejemplo

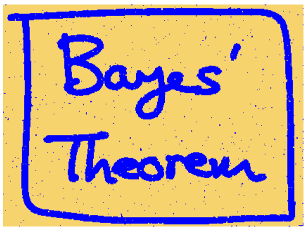
(Bishop. Pattern Recognition and Machine Learning. 2006)

- A partir de la distribución conjunta:

$$P(x_1, \ldots, x_D, y_1, \ldots, y_D) = \frac{1}{Z} \exp\left(\beta \sum_{i,j} x_i x_j + \nu \sum_i x_i y_i\right)$$

- el objetivo es: $(\hat{x}_1, \ldots, \hat{x}_D) = \underset{x_1,\ldots,x_D}{\arg\max} P(x_1, \ldots, x_D \mid y_1, \ldots, y_D)$

Esto es:



Inferencia con campos de Markov aleatorios

- En cadenas: Algoritmo adelante-atrás (”Backward-Forward algorithm”)
- En árboles: Algoritmo suma-producto
- En grafos generales: Algoritmo de árbol de unión (“Junction tree algorithms”), algoritmo suma-producto (“Loopy belief propagation”)

Index

Aprendizaje de redes bayesianas

- Dada la estructura, aprender las distribuciones de probabilidad a partir de un conjunto de entrenamiento.
  - Métodos basados en la maximización de la verosimilitud (algoritmo EM -T3-).
  - Aprendizaje bayesiano.
- Aprender la estructura a partir de un conjunto de entrenamiento.
  - Un problema de selección de modelos: búsqueda en el espacio de grafos.

Algunos toolkits

- BNT
  https://code.google.com/p/bnt
- GMTK
  http://melodi.ee.washington.edu/~bilmes/gmtk/
- GraphLab
  http://graphlab.org/toolkits/graphical-models/
- PMTK3 probabilistic modeling toolkit for Matlab/Octave
  https://github.com/probml/pmtk
- Software Packages for Graphical Models
  http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html

Index

Bibliografía

Christopher M. Bishop: *"Pattern Recognition and Machine Learning"*. Springer, 2006.

Notación

- $P(x)$: probabilidad de x
- $P(x, y)$: probabilidad conjunta de x e y
- $P(x \mid y)$: probabilidad condicional de x dado y
- Para un conjunto de variables V_C en un clique C, $\psi_C(V_C) = \exp(-E(V_C))$ es una función potential donde $E(V_C)$ es una función de energía.
- Una función de energía lineal: $E(V_C) = -\sum_k \theta_{C,k}\, f_{C,k}(V_C)$ donde $f_{C,k}$ son determinadas funciones que obtienen características del clique C.