2020-2021

Aprendizaje Automático

6. Modelos gráficos

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, 2020

Index

Index

Modelos gráficos (MG)

- Los MGs y concretamente las redes bayesianas fundamentan la aproximación probabilística a los Sistemas Inteligentes. Uno de los más famosos impulsores fue Judea Pearl ganador del "ACM A.M. Turing Award" en 2011. Los MGs también se conocen como modelos probabilísticos estructurados.

- Concepto: Representación compacta de distribuciones de probabilidad conjunta mediante grafos 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: deducir distribuciones de probabilidad a partir de otras dadas.
  - Aprendizaje: obtener el modelo probabilístico a partir de observaciones.

- 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, ...

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

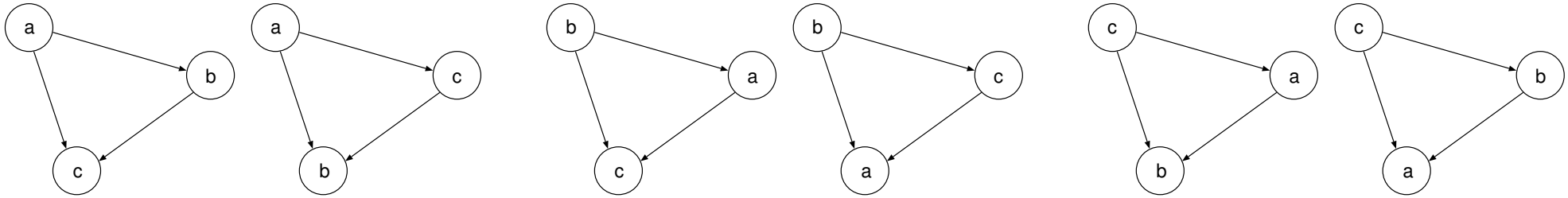
Regla de Bayes $P(y \mid x)\, P(x) = P(y)\, P(x \mid y)$

Factorización de distribuciones conjuntas

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

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

Cada factorización se puede representar mediante un grafo dirigido acíclico:



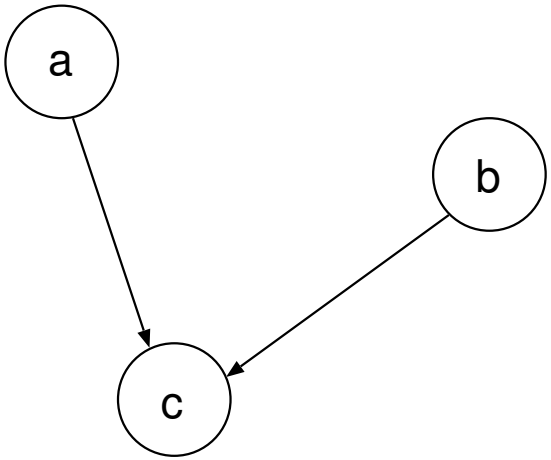
Index

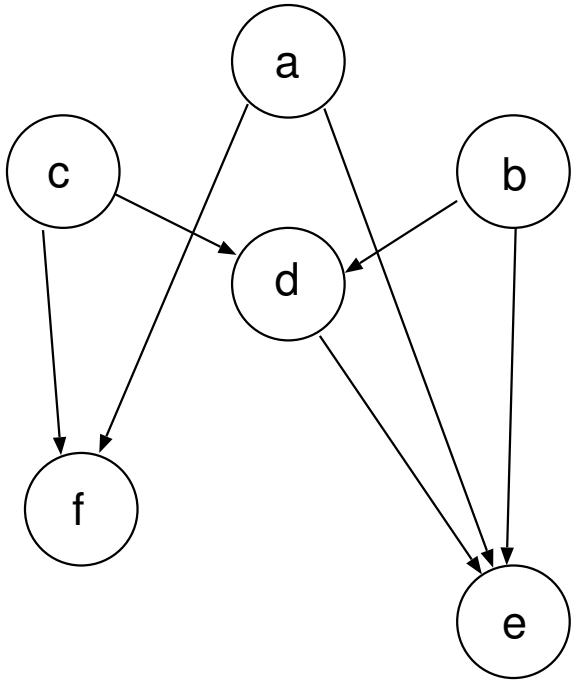
Redes bayesianas: ejemplos

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:

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

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

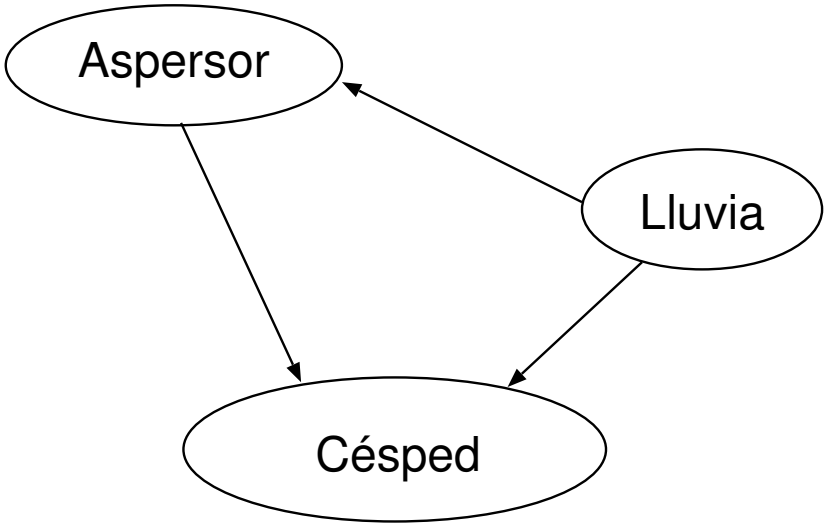




Redes bayesianas: un ejemplo detallado

$P(A \mid L)$

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



$P(L)$

Lluvia (L) n	s
0.8	0.2

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

$P(C \mid A, L)$

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

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

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

Un ejemplo detallado (cont.)

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

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

- El césped está mojado. ¿Cuál es la mejor predicción: llueve o no llueve?

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

Ejercicio:

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

b) Llueve y el cesped está mojado.

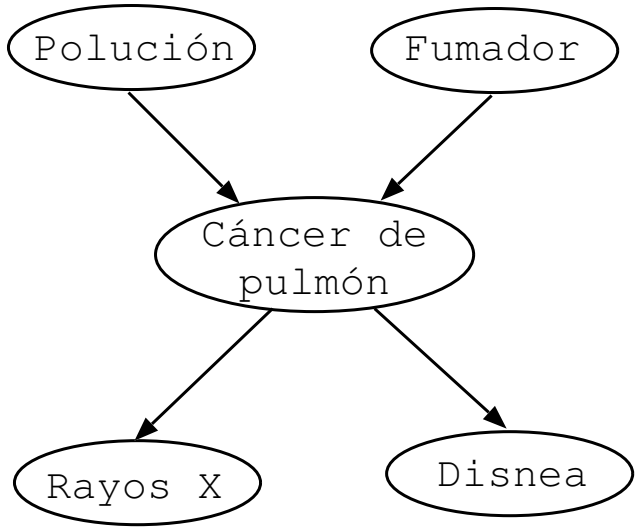
¿Cuál es la mejor predicción sobre el estado del aspersor?

Redes bayesianas: otro ejemplo

$P(P)$

Polución (P)

b	a
0.9	0.1



$P(F)$

Fumador (F)

n	s
0.7	0.3

$P(X \mid C)$

Cáncer (C)	Rayos X (X) n	d	p
n	0.80	0.10	0.10
p	0.10	0.20	0.70

$P(D \mid C)$

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

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

$P(C \mid P, F)$

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

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:

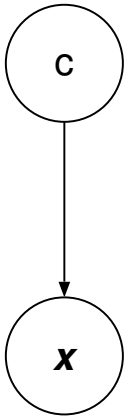
- Los nodos representan:
  - variables aleatorias (discretas o contínuas)
  - distribuciones de probabilidad condicional para cada variable x_i dados los valores de las variables asociadas a los nodos padres $a(x_i)$
- Los arcos representan dependencias entre las variables

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 $c \in \{1, \ldots, C\}$):



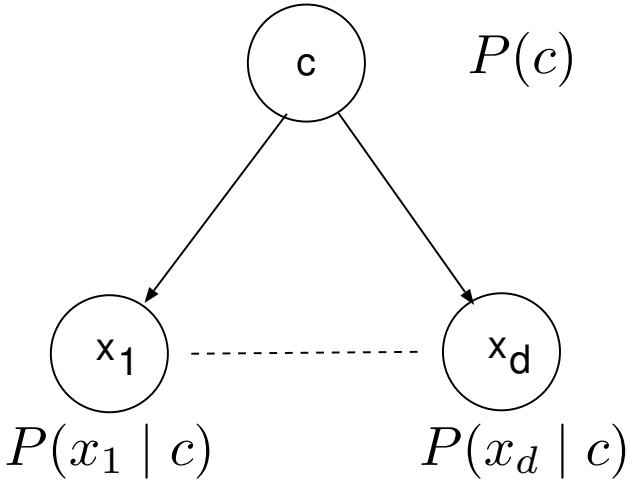
$P(c)$

$P(\boldsymbol{x} \mid c)$

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

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

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

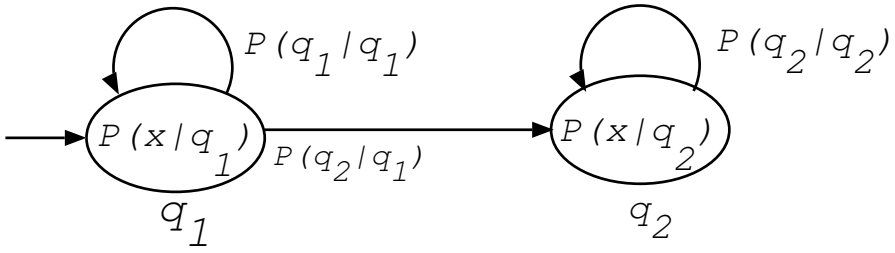


$P(c)$

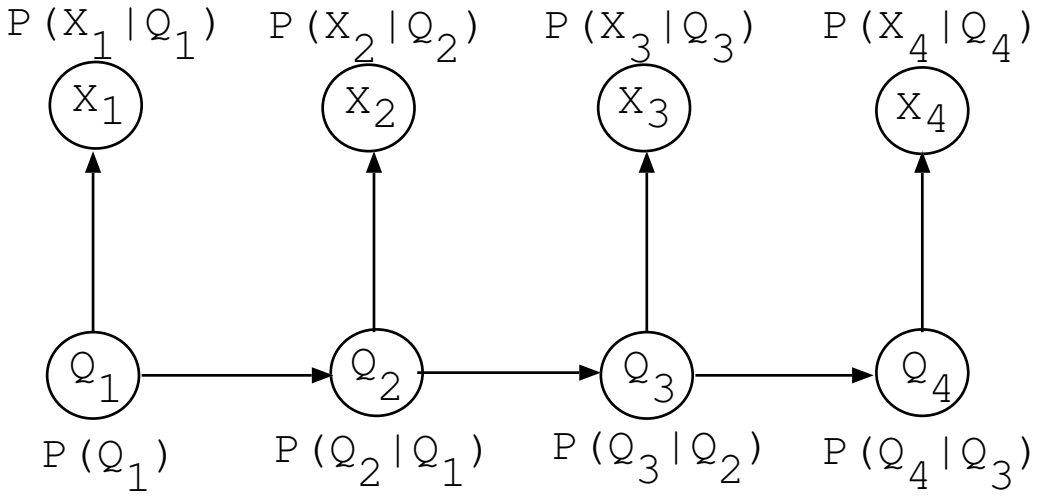
$P(x_1 \mid c)$ $P(x_d \mid c)$

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

Otro ejemplo de red bayesiana: modelo oculto de Markov



$$\Sigma = \{a, b, c\},\ x \in \Sigma$$
$$\mathcal{Q} = \{q_1, q_2\}$$



Las variables aleatorias Q_i toman valores en $\mathcal{Q}$ y las X_i en Σ, $1 \leq i \leq 4$.

Probabilidad conjunta de la red bayesiana: $P(X_1\ X_2\ X_3\ X_4, Q_1\ Q_2\ Q_3\ Q_4) =$

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

Probabilidad de generar la cadena "$aabc$": $P(X_1 = a, X_2 = a, X_3 = b, X_4 = c) =$

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

O sea, la suma de probabilidades de generar "$aabc$" mediante todas las secuencias de 4 estados.

Index

Independencia condicional

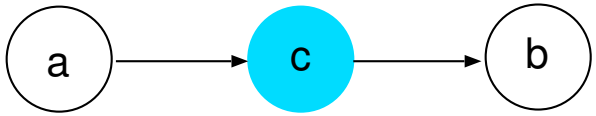
Se dice que a es condicionalmente independiente de b dado c (o también que a está D-separado de b por c, y se denota como: $a \perp\!\!\!\perp b \mid c$) si:

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

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

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

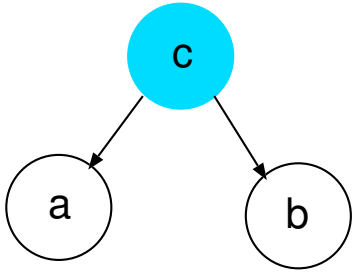
Reglas de independencia condicional e incondicional



Dirección causal: $a \perp\!\!\!\perp b \mid c, \;\; a \not\perp\!\!\!\perp b \mid \emptyset$

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

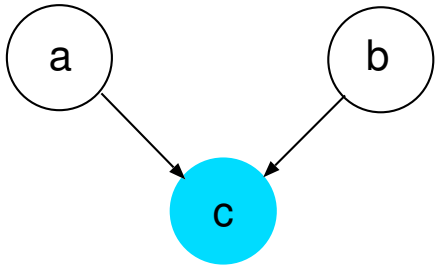
pero en general $P(a, b) \neq P(a)P(b)$



Causa común: $a \perp\!\!\!\perp b \mid c, \;\; a \not\perp\!\!\!\perp b \mid \emptyset$

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

pero en general $P(a, b) \neq P(a)P(b)$



Estructura en V: $a \not\perp\!\!\!\perp b \mid c, \;\; a \perp\!\!\!\perp b \mid \emptyset$

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

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

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 otras variables) y sin importar los valores del resto de las variables f:

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

- El objetivo es calcular eficientemente $P(x,e)$ y $P(e)$

  Ejemplo: Calcular $P(x_3)$ a partir de una distribución conjunta dada por:

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

  Supongamos que cada $x_i,\ i = 1,2,3,4$ puede tomar n valores:

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

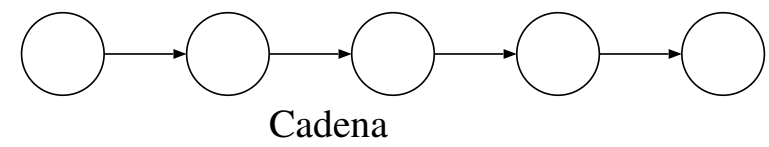
- Diagnóstico: ¿Cuál es la probabilidad de que una determinada enfermedad sea un diagnóstico correcto dados algunos síntomas?

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

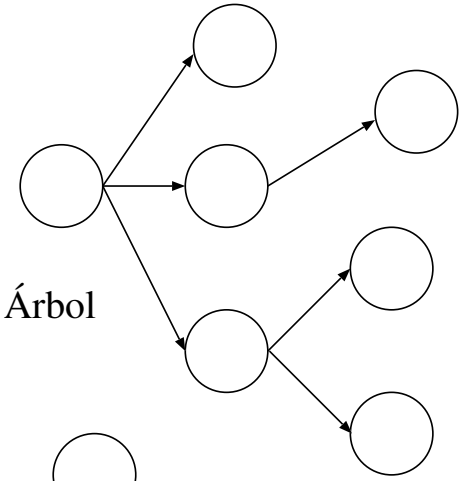
Tipos de redes bayesianas

La estructura de una Red Bayesiana puede permitir ciertas factorizaciones sistemáticas en los cálculos asociados a la inferencia.

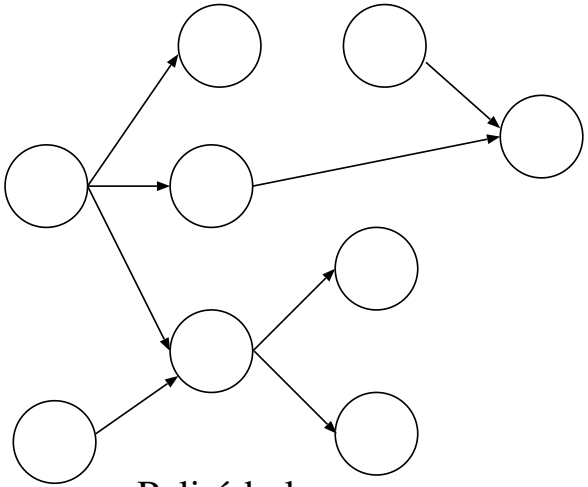
Dos tipos de estructura admiten factorizaciones eficientes: *cadena* y *(poli-)árbol*.



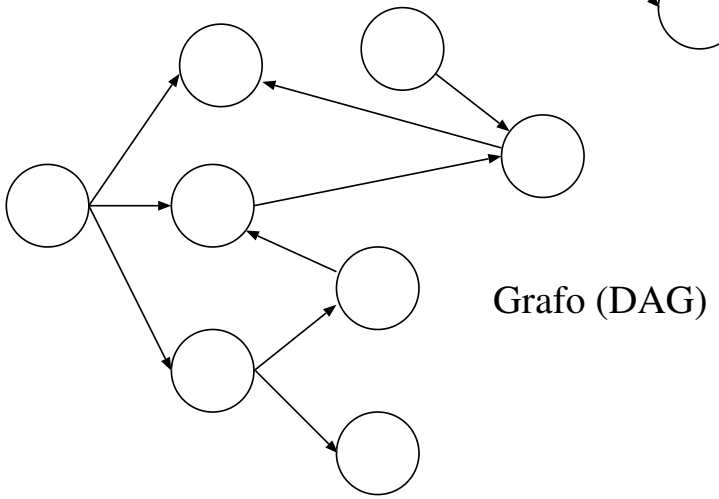
Cadena



Árbol

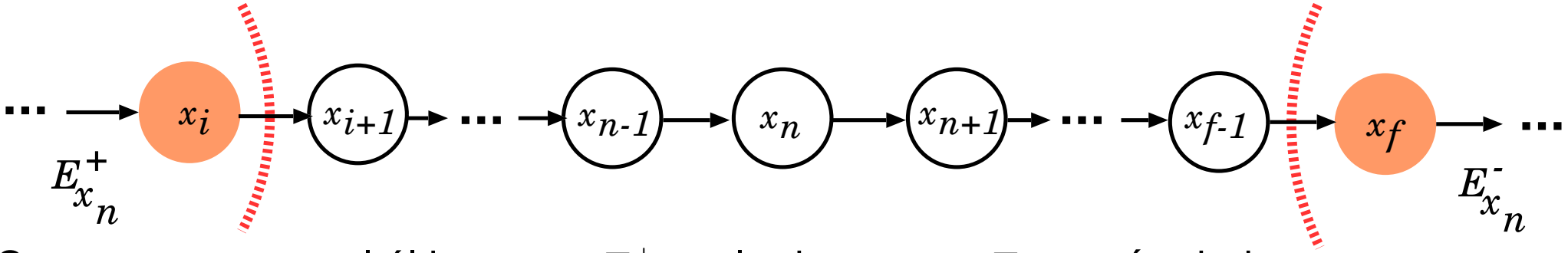


Poli-árbol



Grafo (DAG)

Inferencia en una cadena

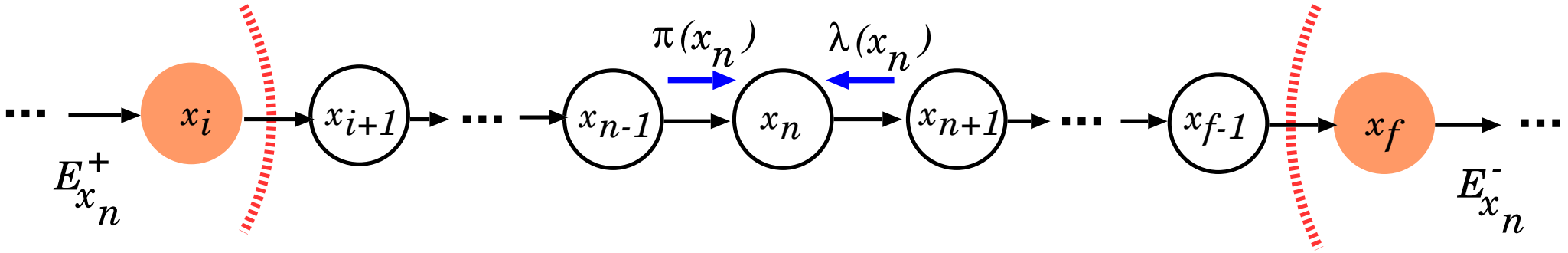


Supongamos que el último $x_i \in E^+_{x_n}$ y el primer $x_f \in E^-_{x_n}$ están dados:

$$
\begin{aligned}
P(x_n \mid x_i, x_f) &= \frac{P(x_n, x_i, x_f)}{P(x_i, x_f)} \\
&= \frac{P(x_n)\, P(x_i \mid x_n)\, P(x_f \mid x_n, \cancel{x_i})}{P(x_i, x_f)} \qquad \text{(Indep. cond.: } x_f \perp\!\!\!\perp x_i \mid x_n\text{)} \\
&= \frac{\cancel{P(x_n)}\, P(x_i)\, P(x_n \mid x_i)\, P(x_f \mid x_n)}{\cancel{P(x_n)}\, P(x_i, x_f)} \qquad \text{(Regla de Bayes)} \\
&= \alpha\, P(x_n \mid x_i)\, P(x_f \mid x_n) \qquad (\alpha = P(x_i)/P(x_i, x_f))
\end{aligned}
$$

- 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
Propagación de creencias ("Belief propagation")



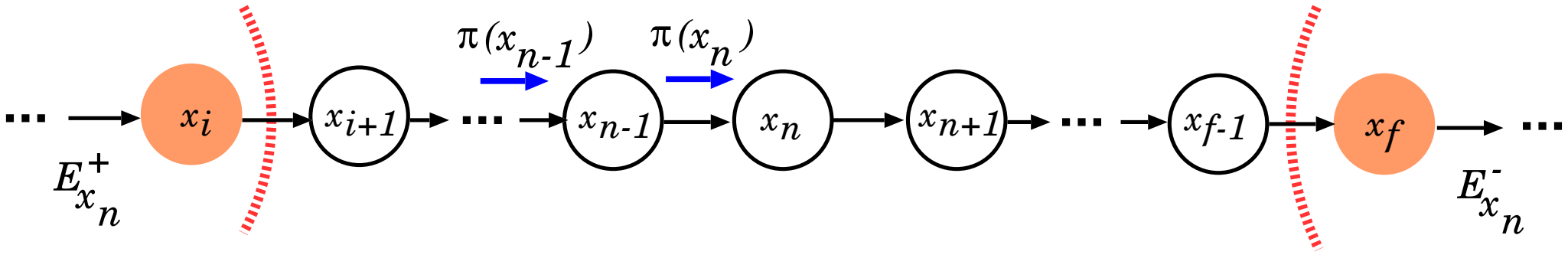
$$P(x_n \mid x_i, x_f) \;=\; \alpha\, P(x_n \mid x_i)\, P(x_f \mid x_n) \;\stackrel{\text{def}}{=}\; \alpha\, \pi(x_n)\, \lambda(x_n)$$

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

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

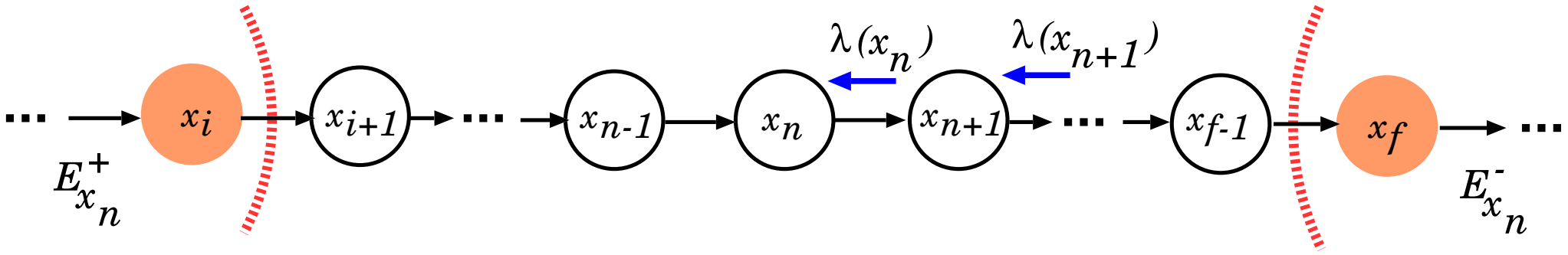
- Ejercicio: Obtener una expresión para $P(x_n)$

Inferencia en una cadena (derivación I)



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

Inferencia en una cadena (derivación II)



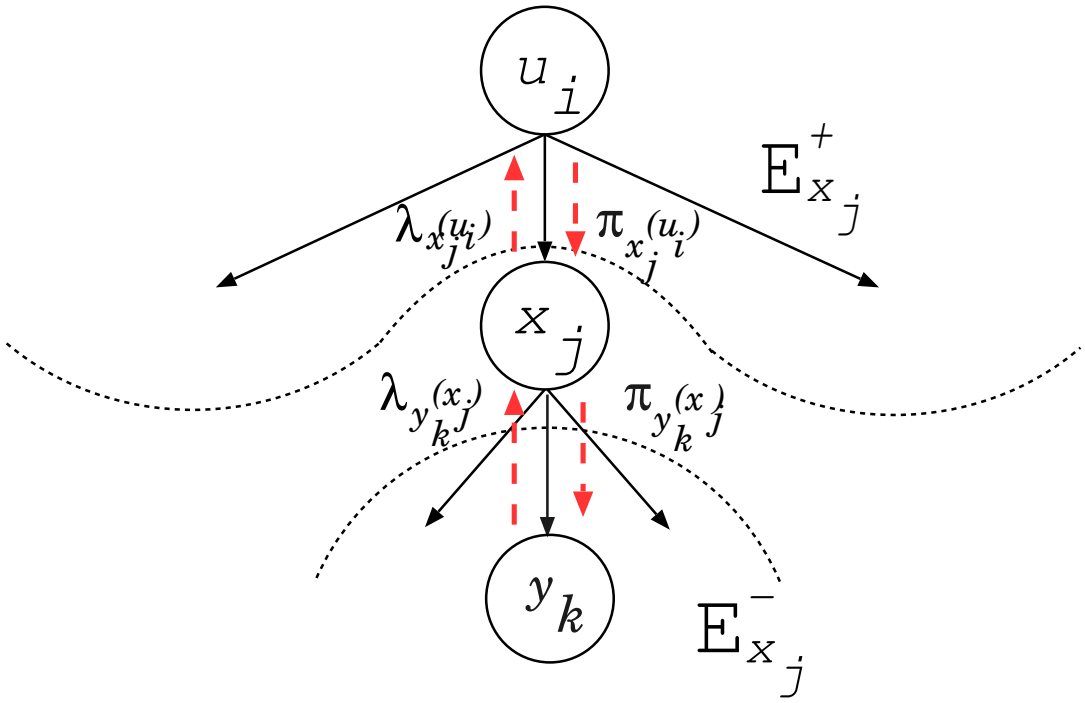
$$\lambda(x_n) = P(x_f \mid x_n) = \sum_{x_{n+1}} P(x_f, x_{n+1} \mid x_n)$$

$$= \sum_{x_{n+1}} P(x_{n+1} \mid x_n)\, P(x_f \mid \cancel{x_n}, x_{n+1}) = \sum_{x_{n+1}} P(x_{n+1} \mid x_n)\, P(x_f \mid x_{n+1})$$

$$= \sum_{x_{n+1}} P(x_{n+1} \mid x_n)\, \lambda(x_{n+1})$$

$$\lambda(x_f) = 1$$

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 nodos 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 muestreo.

Index

Campos de Markov aleatorios

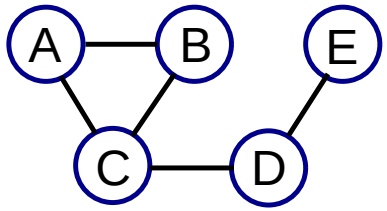
Un campo de Markov aleatorio es un modelo gráfico en la que se asume un modelo simplificado de independencia condicional. Se define como:

- Un conjunto de variables aleatorias $V = \{X_1, \ldots, X_D\}$
- Un grafo no-dirigido $R = (V, E)$
- El conjunto $\mathcal{Q}$ de todos los *cliqués (máximos)*[†] de R
- Una distribución de probabilidad conjunta factorizada como:

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

donde V_C es el subconjunto de variables del cliqué C y $\psi_C : \mathcal{Q} \to \mathbb{R}^{>0}$ es una función llamada función potential y Z es un factor de normalización.

Ejemplo (3 cliqués máximos, $V_1 = \{A, B, C\},\ V_2 = \{C, D\},\ V_3 = \{D, E\}$):



$$P(A, B, C, D, E) = \frac{1}{Z} \psi_1(A, B, C) \cdot \psi_2(C, D) \cdot \psi_3(D, E)$$

[†]Un cliqué es un subgrafo totalmente conectado. Es *máximo* si no es subgrafo de algún otro cliqué.

Campos de Markov aleatorios: potenciales exponenciales

Si las funciones de potencial son de la familia exponencial:

$$\begin{aligned} P(x_1, \ldots, x_D) &= \frac{1}{Z} \prod_{C \in \mathcal{Q}} \psi_C(V_C) \\ &= \frac{1}{Z} \prod_{C \in \mathcal{Q}} \exp\left(-E_C(V_C)\right) \\ &= \frac{1}{Z} \exp\Big(- \sum_{C \in \mathcal{Q}} E_C(V_C) \Big) \end{aligned}$$

donde $E_C : \mathcal{Q} \to \mathbb{R}$ es una función llamada función de energía.

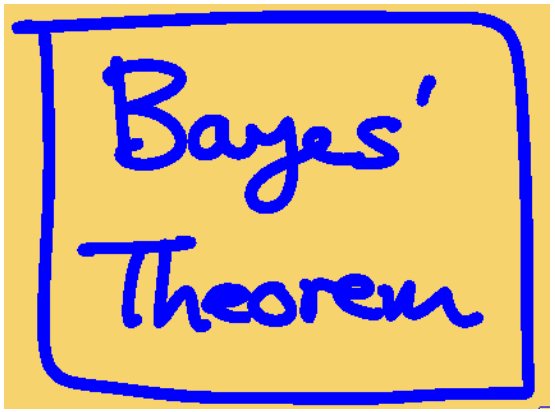
Un tipo de función de energía simple puede definirse mediante funciones lineales generalizadas:

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

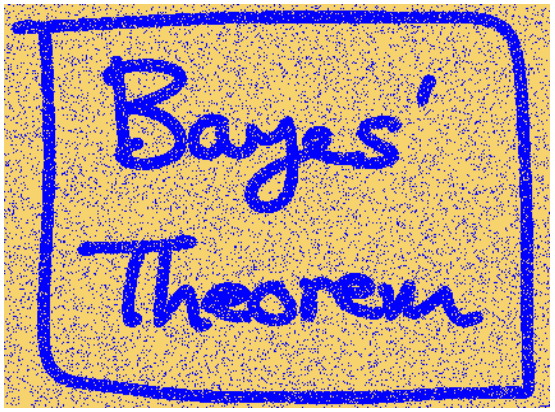
Un ejemplo

(Bishop. Pattern Recognition and Machine Learning. 2006)

Imagen binaria original: $\boldsymbol{x}$,
$x_i \in \{-1, +1\},\ 1 \leq i \leq D$
($D =$ número de píxeles de la imagen)



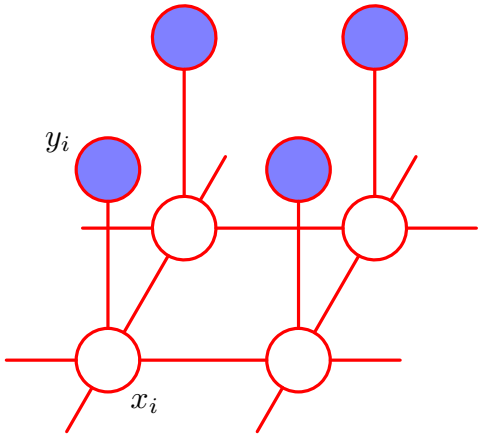
La imagen corrupta: $\boldsymbol{y}$,
$y_i \in \{-1, +1\}, 1 \leq i \leq D$
(se han alterado 10% de los píxeles)



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 son píxeles vecinos; es decir, si $i \in N(j), j \in N(i)$.
- Cliques máximos: $C_i = (x_i, y_i)\ \forall i; \quad C_{ij} = (x_i, x_j)\ \forall i, \forall j \in N(i)$
- Función de energía:

$$\left.\begin{array}{l} E_C(V_{C_{ij}}) = -\beta\ x_i\ x_j \\ E_C(V_{C_i}) = -\nu\ x_i\ y_i \end{array}\right\} \rightarrow \sum_{C \in \mathcal{Q}} E_C(V_C) = -\beta \sum_{i,j} x_i\ x_j - \nu \sum_i x_i\ y_i$$

- 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:

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

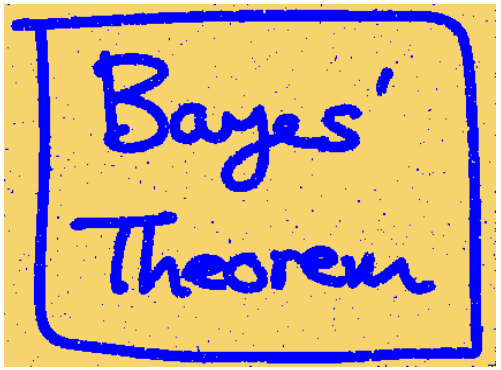
- Inferencia:

$$P(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{P(\boldsymbol{x}, \boldsymbol{y})}{P(\boldsymbol{y})} = \frac{\exp\left(\beta \sum_{i,j} x_i x_j + \nu \sum_i x_i y_i\right)}{\sum_{x'_1,\ldots,x'_D} \exp\left(\beta \sum_{i,j} x'_i x'_j + \nu \sum_i x'_i y_i\right)}$$

- Explicación más probable:

$$\hat{\boldsymbol{x}} \equiv (\hat{x}_1, \ldots, \hat{x}_D) = \arg\max_{\boldsymbol{x}} P(\boldsymbol{x} \mid \boldsymbol{y}) = \arg\max_{x_1,\ldots,x_D} \left(\beta \sum_{i,j} x_i x_j + \nu \sum_i x_i y_i\right)$$

Esto es, $\hat{\boldsymbol{x}} =$



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

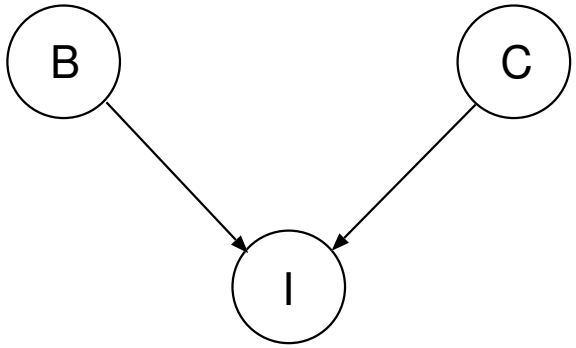
- Dada la estructura, aprender las distribuciones de probabilidad a partir de un conjunto de entrenamiento.
  - Métodos basados en la maximización de la verosimilitud (algoritmo EM -T3-).
  - Aprendizaje bayesiano.
- Aprender la estructura a partir de un conjunto de entrenamiento.
  - Un problema de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

Aprendizaje en redes bayesianas: ejemplo

- 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 $I = 0$)

Por simplicidad se asume que las distribuciones asociadas a las variables I y C son fijas y conocidas y hay que estimar la distribución de B a partir de observaciones.

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



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

$P(B)$ viene dada por dos valores de probabilidad, $p_0 = P(B=0)$, $p_1 = P(B=1)$. Por tanto los parámetros a estimar son:

$$\boldsymbol{\Theta} = (p_0, p_1), \quad p_0 + p_1 = 1$$

Aprendizaje con observaciones completas

Sea $S = \{(b_1, c_1, i_1), \ldots, (b_N, c_N, i_N),\}$ un conjunto de observaciones de entrenamiento. Sean N_0 y $N_1 = N - N_0$ los números de observaciones en las que $b = 0$ y $b = 1$, respectivamente. Se asume que S está ordenado de forma que $b_n = 0, \quad 1 \leq n \leq N_0$ y $b_m = 1, \; N_{N_0+1} \leq m \leq N$. La log-verosimilitud es:

$$\begin{aligned} L_S(p_0, p_1) &= \sum_{n=1}^{N} \log P(B = b_n, C = c_n, I = i_n) \\ &= \sum_{n=1}^{N} \log(P(B = b_n) P(C = c_n) P(I = i_n \mid B = b_n, C = c_n)) \\ &= \sum_{n=1}^{N_0} \log p_0 + \sum_{m=N_0+1}^{N} \log p_1 + K \;=\; N_0 \log p_0 + N_1 \log p_1 + K \end{aligned}$$

donde K incluye los términos independientes de p_0, p_1. Maximización de L_S mediante multiplicadores de Lagrange, como en el ejemplo de estimación de probs. a priori del Tema 3 $\cdots \longrightarrow$

$$\hat{p}_0 = \frac{N_0}{N}, \qquad \hat{p}_1 = \frac{N_1}{N}$$

Ejemplos:

$$\begin{aligned} S = \{(0,1,1),(1,1,0)\} &\rightarrow & \hat{p}_0 = \tfrac{1}{2}, \quad \hat{p}_1 = \tfrac{1}{2} \\ S = \{(0,0,1),(1,1,0)\} &\rightarrow & \hat{p}_0 = \tfrac{1}{2}, \quad \hat{p}_1 = \tfrac{1}{2} \\ S = \{(0,1,1),(0,0,1),(1,1,0)\} &\rightarrow & \hat{p}_0 = \tfrac{2}{3}, \quad \hat{p}_1 = \tfrac{1}{3} \end{aligned}$$

Algoritmo esperanza-maximización (EM) (Recordatorio)

Dada la muestra $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ y variables latentes $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$.

- Inicialización: $t = 1$, $\boldsymbol{\Theta}(1) =$ arbitrario.
- Iteración hasta la convergencia:
  - Paso E: A partir de $\boldsymbol{\Theta}(t)$ y para todo $1 \leq n \leq N$,

$$\hat{q}_n(\boldsymbol{z}_n) \triangleq P(\boldsymbol{z}_n \mid \boldsymbol{x}_n; \boldsymbol{\Theta}(t))$$

  - Paso M: Maximización de $Q(\boldsymbol{\Theta}, \hat{\boldsymbol{q}})$ con respecto a $\boldsymbol{\Theta}$

$$\boldsymbol{\Theta}(t+1) = \arg\max_{\boldsymbol{\Theta}} Q'(\boldsymbol{\Theta}, \hat{\boldsymbol{q}}) = \arg\max_{\boldsymbol{\Theta}} \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} \hat{q}_n(\boldsymbol{z}_n) \log P(\boldsymbol{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\Theta})$$

  - $t = t + 1$

Algoritmo esperanza-maximización (EM)

En el ejemplo, se dispone de observaciones del estado del depósito de combustible y de lo que marque el indicador del estado de batería, pero *no* del verdadero estado de la batería, la muestra incompleta serán los datos observables, $S = \{(c_1, i_1), \ldots, (c_N, i_N)\}$, ($\boldsymbol{x} = (c, i)$), las variables latentes, $\{b_1, \ldots, b_N\}$ ($\boldsymbol{z} = b$) y los parámetros p_0 y p_1 ($\boldsymbol{\Theta} = (p_0, p_1)$).

- Inicialización: $t = 1$, $p_0(1) =$ arbitrario y $p_1(1) = 1 - p_0(1)$.
- Iteración hasta la convergencia
  - Paso E. A partir de $p_0(t)$ y $p_1(t)$ y cada muestra $1 \leq n \leq N$.

$$\hat{q}_n(b) \triangleq P(B = b \mid C = c_n, I = i_n; p_0(t), p_1(t)) \quad b \in \{0, 1\}$$

  - Paso M.

$$\begin{aligned}(p_0(t+1), p_1(t+1)) &= \arg\max_{p_0, p_1} Q'(p_0, p_1, \hat{\boldsymbol{q}}) \\ &= \arg\max_{p_0, p_1} \sum_{n=1}^{N} \sum_{b=0}^{1} \hat{q}_n(b) \log P(C = c_n, I = i_n, B = b \mid p_0, p_1)\end{aligned}$$

  - $t = t + 1$

Aprendizaje EM con observaciones incompletas: Paso E

En el paso E hay que calcular, para cada observación incompleta, la distribución de probabilidad de la variable oculta o latente (B) dados los valores conocidos de las otras dos variables (C e I) y de los parámetros obtenido en la iteración anterior t ($p_0(t), p_1(t)$).

$$\begin{aligned}
\hat{q}_n(b) &= P(B{=}b \mid C = c_n, I = i_n; p_0(t), p_1(t)) \\
&= \frac{P(B{=}b, C = c_n, I = i_n; p_0(t), p_1(t))}{P(C = c_n, I = i_n; p_0(t), p_1(t))} \\
&= \frac{P(B{=}b; p_0(t), p_1(t))\cancel{P(C{=}c_n)}\,P(I{=}i_n|B{=}b, C{=}c_n)}{P(B{=}0)\cancel{P(C{=}c_n)}\,P(I{=}i_n|B{=}0, C{=}c_n) + P(B{=}1)\cancel{P(C{=}c_n)}\,P(I{=}i_n|B{=}1, C{=}c_n)} \\
&= \frac{p_b(t)\; k_{bn}}{p_0(t)\; k_{0n} + p_1(t)\; k_{1n}} \quad \text{para} \quad b \in \{0, 1\}
\end{aligned}$$

Donde: $k_{0n} = P(I{=}i_n \mid B{=}0, C{=}c_n)$ y $k_{1n} = P(I{=}i_n \mid B{=}1, C{=}c_n)$

Aprendizaje EM con observaciones incompletas: paso M

En el paso M hay que maximizar $Q'(p_0, p_1, \hat{\boldsymbol{q}})$, definida en base a las esperanzas calculadas en el paso E.

$$\begin{aligned} Q'(p_0, p_1, \hat{\boldsymbol{q}}) &= \sum_{n=1}^{N} \sum_{b=0}^{1} \hat{q}_n(b) \ \log P(B = b, C = c_n, I = i_n; p_0, p_1) \\ &= \sum_{n=1}^{N} \sum_{b=0}^{1} \hat{q}_n(b) \log(P(B = b; p_0, p_1) P(C = c_n) P(I = i_n \mid B = b, C = c_n)) \\ &= \sum_{n=1}^{N} \hat{q}_n(0) \log p_0 + \hat{q}_n(1) \log p_1 + K \ = \ \bar{q}_0 \ \log p_0 + \bar{q}_1 \ \log p_1 + K \end{aligned}$$

Donde,

$$K = \sum_{n=1}^{N} \sum_{b=0}^{1} \hat{q}_n(b) \log(P(C = c_n) P(I = i_n \mid B = b, C = c_n))$$

$$\bar{q}_0 = \sum_{n=1}^{N} \hat{q}_n(0) \ \text{ y } \ \bar{q}_1 = \sum_{n=1}^{N} \hat{q}_n(1)$$

.

Aprendizaje EM con observaciones incompletas: paso M

Maximizar Q': similar a estimación de probabilidades a priori por máxima verosimilitud (Tema 3):

- Función de Lagrange:

$$\begin{aligned}\Lambda(p_0, p_1, \beta) &= Q'(p_0, p_1, \hat{\boldsymbol{q}}) + \beta(1 - p_0 - p_1) \\ &= \bar{q}_0 \log p_0 + \bar{q}_1 \log p_1 + K + \beta(1 - p_0 - p_1)\end{aligned}$$

- Función dual: $\dfrac{\partial \Lambda}{\partial p_0} = 0 \Rightarrow p_0^\star = \dfrac{\bar{q}_0}{\beta}; \qquad \dfrac{\partial \Lambda}{\partial p_1} = 0 \Rightarrow p_1^\star = \dfrac{\bar{q}_1}{\beta}$

$$\Lambda_D(\beta) = K' + \beta - (\bar{q}_0 + \bar{q}_1)\log\beta = K' + \beta - N\log\beta$$

$$\text{Donde } K' = K + \bar{q}_0 \log \bar{q}_0 + \bar{q}_1 \log \bar{q}_1$$

- Optimizando la función dual: $\dfrac{d\Lambda_D}{d\beta} = 1 - \dfrac{N}{\beta} = 0 \Rightarrow \beta^\star = N$

- Solución final: $p_0(t+1) \triangleq p_0^\star = \dfrac{\bar{q}_0}{N};\ p_1(t+1) \triangleq p_1^\star = \dfrac{\bar{q}_1}{N}$

Algoritmo esperanza-maximización (EM)

Dado un conjunto de datos obervables $S = \{(c_1, i_1), \ldots, (c_N, i_N)\}$, ($\boldsymbol{x} = (c, i)$), las variables latentes, $\{b_1, \ldots, b_N\}$ ($\boldsymbol{z} = b$) y los parámetros p_0 y p_1 ($\boldsymbol{\Theta} = (p_0, p_1)$)., el algoritmo final queda:

- Inicialización: $t = 1$, $p_0(1) =$ arbitrario y $p_1(1) = 1 - p_0(1)$.
- Iteración hasta la convergencia
  - Paso E. A partir de $p_0(t)$ y $p_1(t)$.
    - Para cada muestra $1 \leq n \leq N$

$$\hat{q}_n(b) = \frac{p_b(t)\ k_{bn}}{p_0(t)\ k_{0n} + p_1(t)\ k_{1n}} \quad \text{para} \quad b \in \{0, 1\}$$

    - Calcular:

$$\bar{q}_0 = \sum_{n=1}^{N} \hat{q}_n(0) \quad \bar{q}_1 = \sum_{n=1}^{N} \hat{q}_n(1)$$

  - Paso M. Calcular: $(p_0(t+1), p_1(t+1)) = \left(\frac{\bar{q}_0}{N}, \frac{\bar{q}_1}{N}\right)$
  - $t = t + 1$

Aprendizaje EM con observaciones incompletas: ejemplos

Sea $S = \{(1,1),(1,0)\}$ $(N = 2)$:

Paso E: $\hat{q}_1(0) = 0.2p_0(t)/(0.2p_0(t) + 0.8p_1(t))$, $\hat{q}_1(1) = 0.8p_0(t)/(0.2p_0(t) + 0.8p_1(t))$,

$\hat{q}_2(0) = 0.8p_0(t)/(0.8p_0(t) + 0.2p_1(t))$, $\hat{q}_2(1) = 0.2p_0(t)/(0.8p_0(t) + 0.2p_1(t))$.

$\bar{q}_0 = \hat{q}_1(0) + \hat{q}_2(0)$, $\bar{q}_1 = \hat{q}_1(1) + \hat{q}_2(1)$.

Paso M: $p_0(t+1) = \dfrac{\bar{q}_0}{N}$, $p_1(t+1) = \dfrac{\bar{q}_1}{N}$

Inicializando con $p_1(1) = 0.9,\ p_0(1) = 0.1$:

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$p_1(t)$	0.90	0.83	0.75	0.68	0.62	0.58	0.55	0.53	0.52	0.51	0.51	0.51	0.50	0.50	0.50	0.50

Si ahora las observaciones son $S = \{(0,1),(1,0)\}$:

t	1	2	3	4	5	10	20	25	30	35	36	37	38	39	40	41
$p_1(t)$	0.90	0.82	0.72	0.61	0.52	0.45	0.28	0.19	0.18	0.18	0.17	0.17	0.17	0.17	0.17	0.17

Algunos toolkits

- BNT
  https://code.google.com/p/bnt
- GraphLab
  https://turi.com/
- 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, $\psi_C(V_C) = \exp\left(-E(V_C)\right)$ 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.