Redes Bayesianas (3)

Carlos Hurtado L.

Depto. de Ciencias de la Computación, Universidad de Chile

Referencia

Tutorial NIPS (Neural Information Processing Systems Conference) 2001: Learning Bayesian Networks from Data. Nir Friedman and Daphne Koller

Redes Bayesianas

Representación compacta de distribución conjunta via aserciones de independencia condicional.

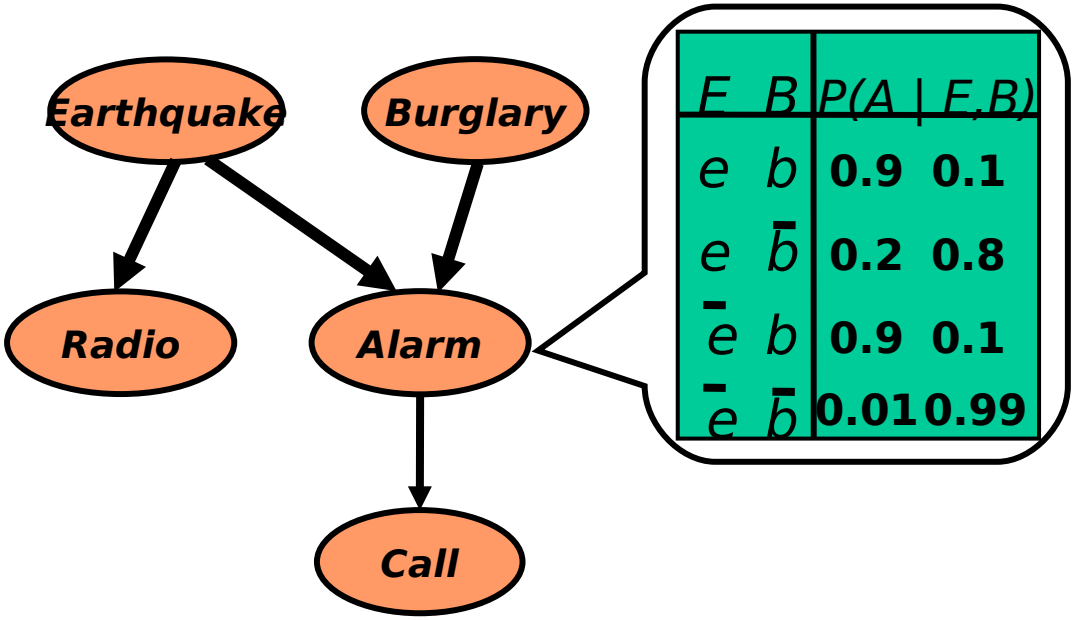


Tabla de Probabilidades Condicionales (TPC)

Las TPCs definen una distribución única en forma "factorizada"

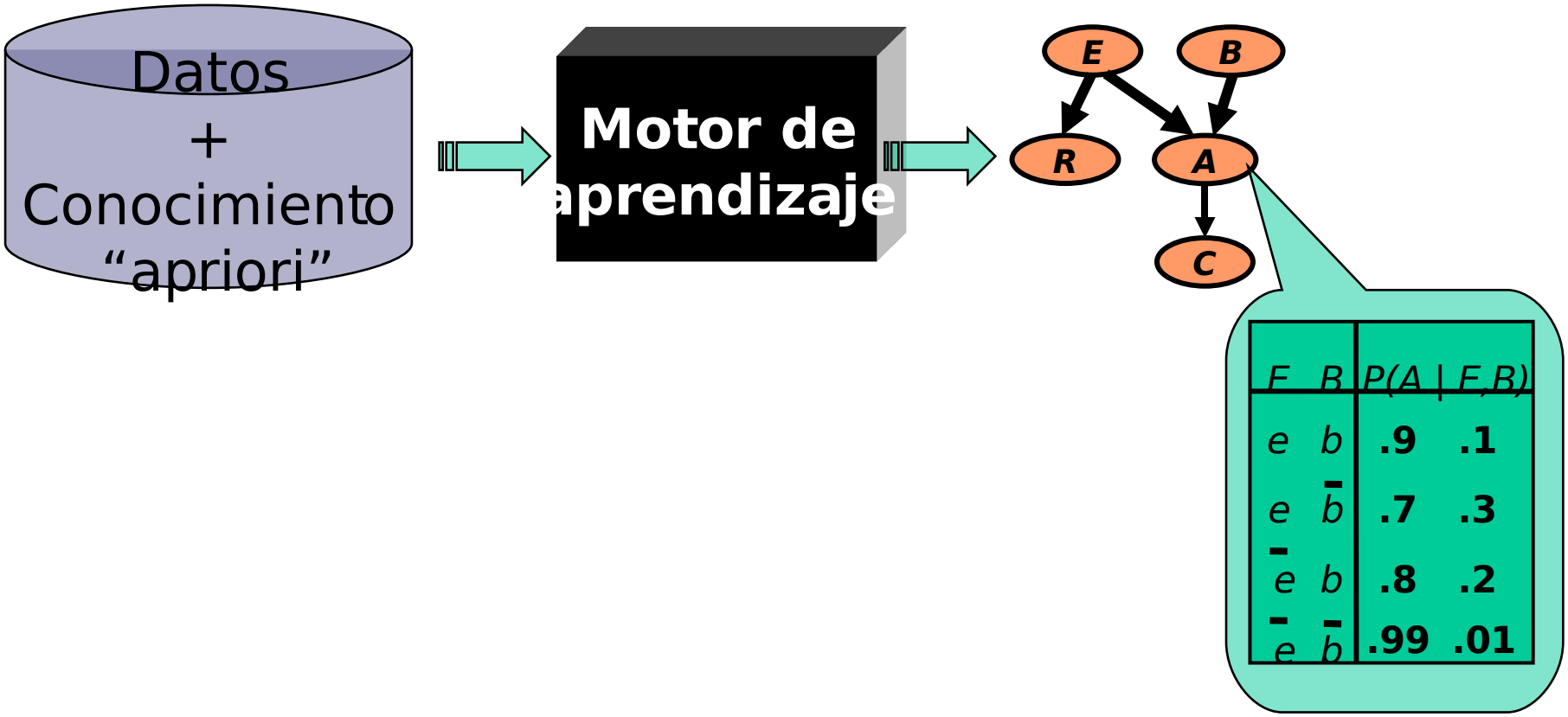
$$P(B,E,A,C,R)=P(B)P(E)P(A|B,E)P(R|E)P(C|A)$$

Construcción de una Red Bayesiana

Veremos tres casos:

- Red y TPCs se definen por expertos
- Red predefinida y TPCs inducidas de los datos
- Red y TPCs inducidas de los datos

Aprendizaje de Redes Bayesianas



Construcción de una Red Bayesiana

Veremos tres casos:

- Red y TPCs se definen por expertos
- Red predefinida y TPCs inducidas de los datos
- Red y TPCs inducidas de los datos

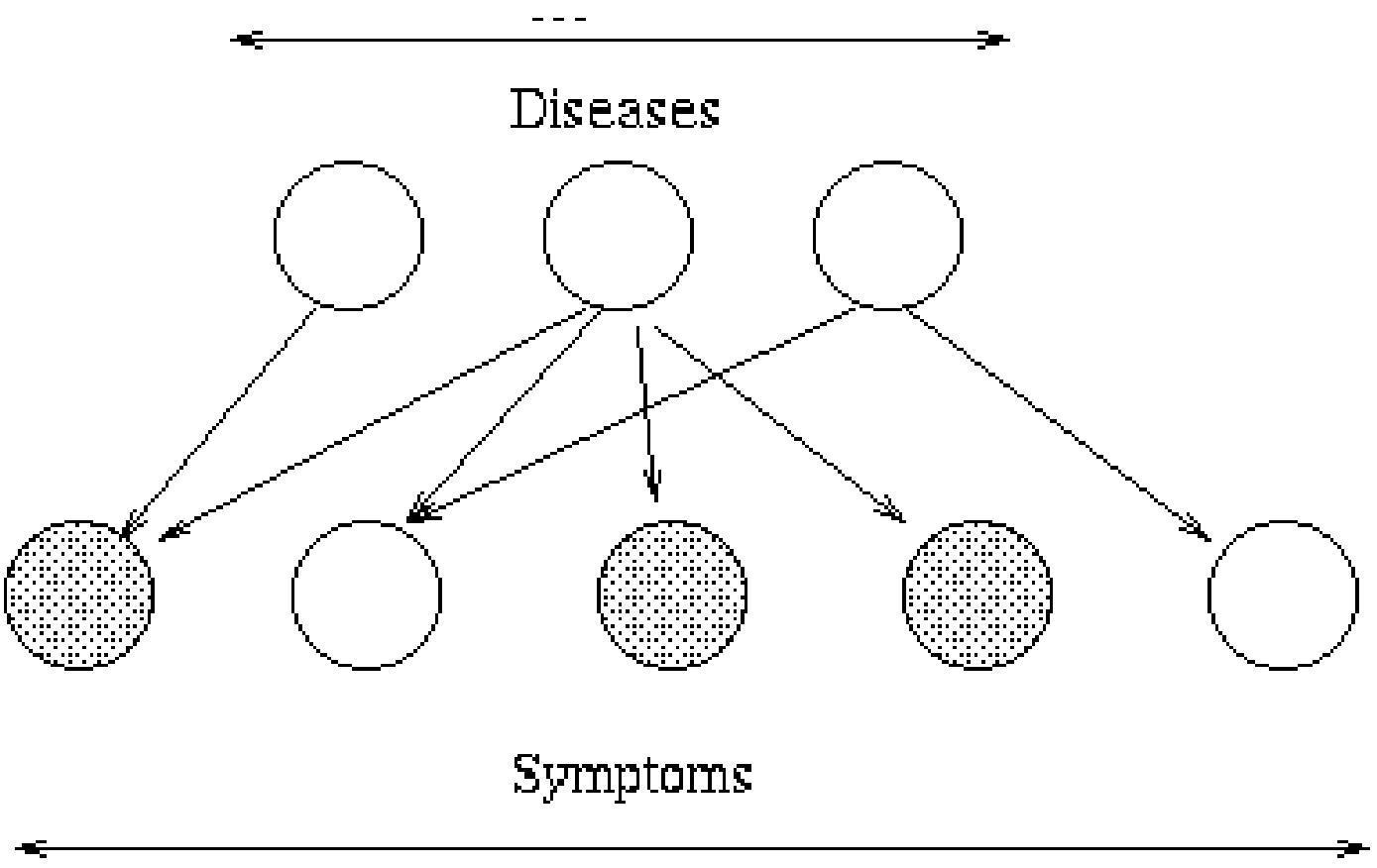
Red y TPCs predefinidas

- La red Bayesiana completa codifica conocimiento “apriori” sobre un sistema
- Ejemplo: Pathfinder

PathFinder

- Diagnóstico de enfermedades del sistema linfático
  - Entrada: observación de 100 variables (síntomas) de una biopsia
  - Salida: Una entre 60 tipos de enfermedades malignas y benignas.
- En la actualidad PathFinder supera a expertos mundiales en el área.
- TPCs contienen 14.000 parámetros aprox.

PathFinder: Estructura de la Red



Construcción de PathFinder

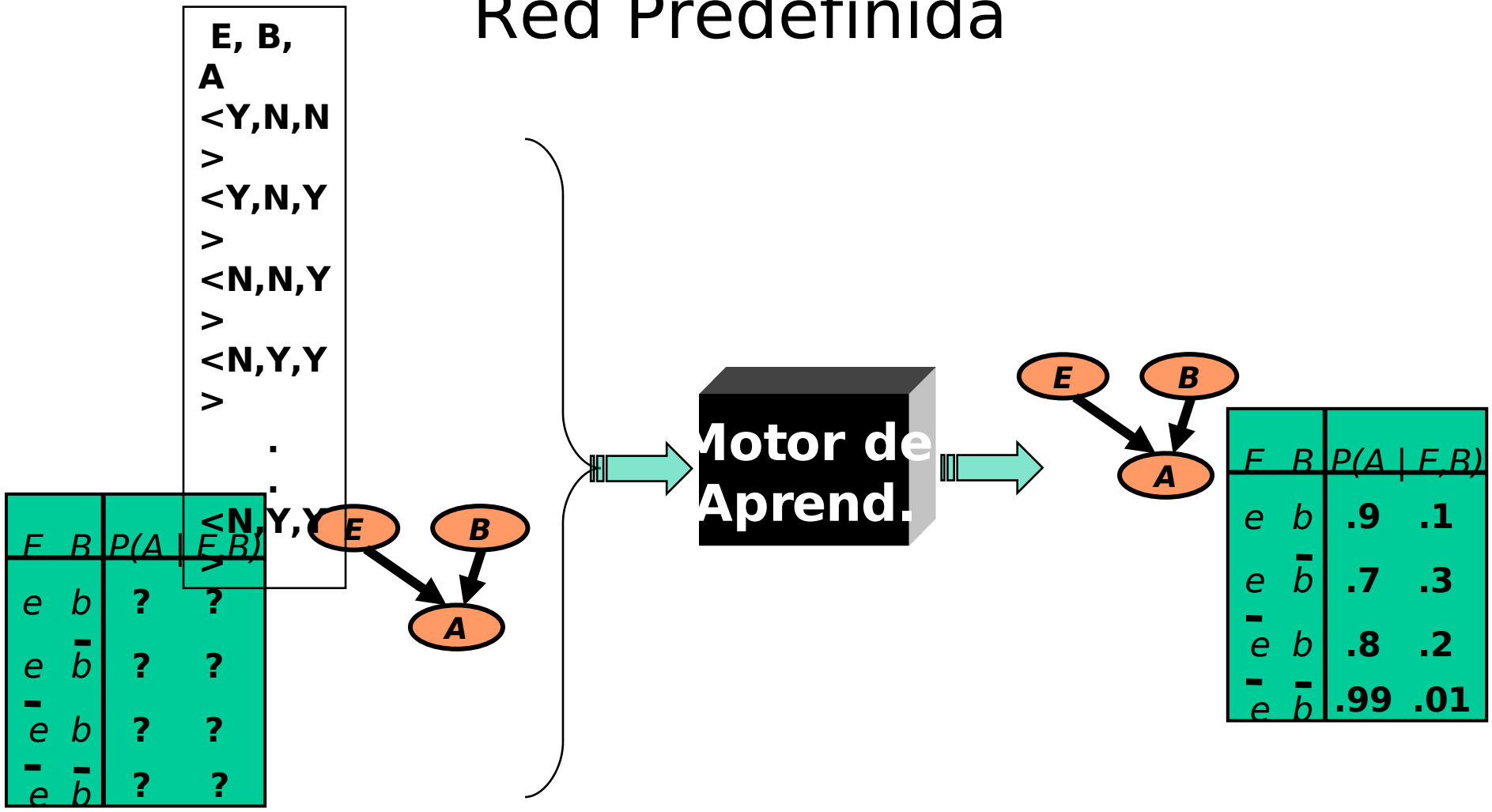
- Tanto la red como las TPCs se consultan a expertos
- Expertos definen conexiones entre enfermedades y síntomas
- Tiempo de construcción:
  - 8 hrs en definir variables
  - 35 hrs para definir estructura de la red
  - 40 hrs para llenar TPCs

Construcción de una Red Bayesiana

Veremos tres casos:

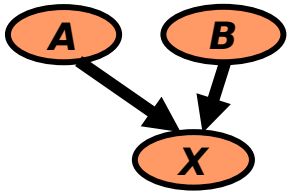
- Red y TPCs se definen por expertos
- Red predefinida y TPCs inducidas de los datos
- Red y TPCs inducidas de los datos

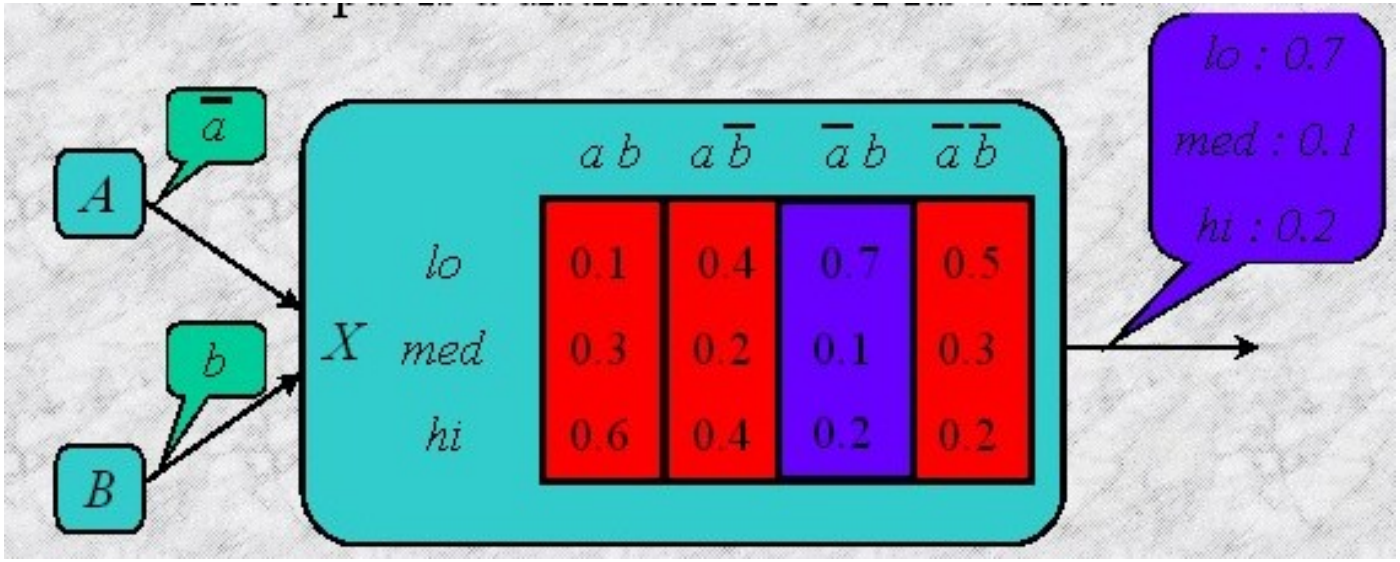
Construcción de Red Bayesiana: Red Predefinida



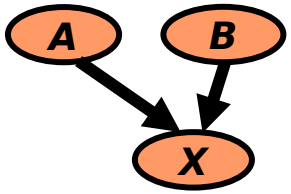
El problema se reduce a estimar las TPCs

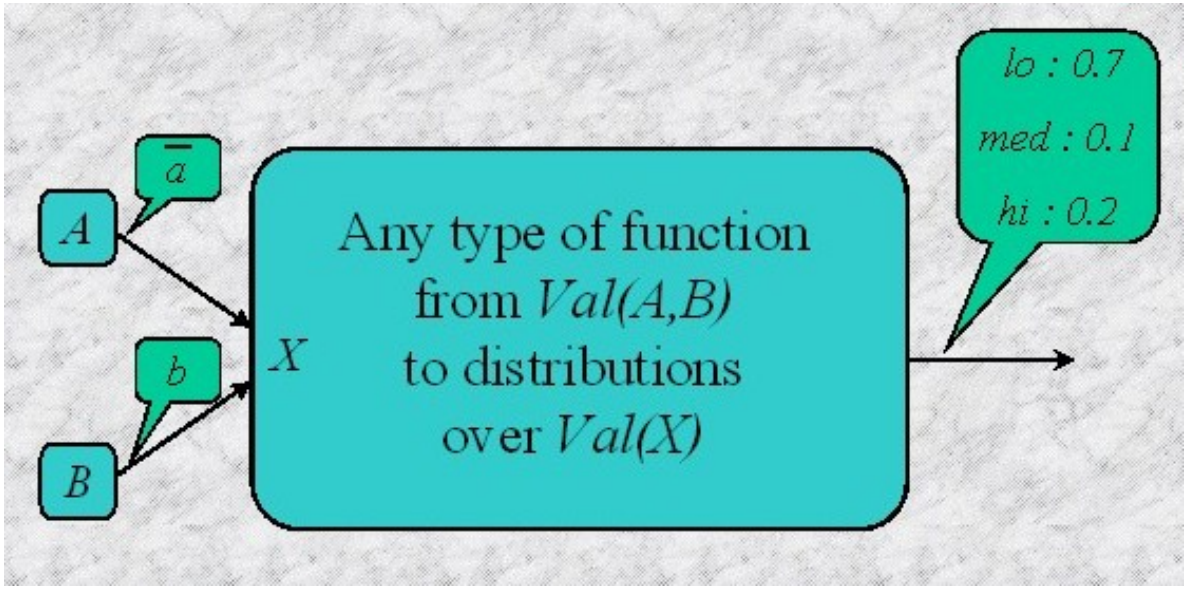
TPCs vistas como Funciones



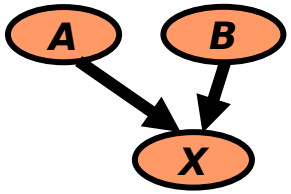


TPCs vistas como Funciones





TPCs: dos casos



- X es variable aleatoria discreta
- X es variable aleatoria continua

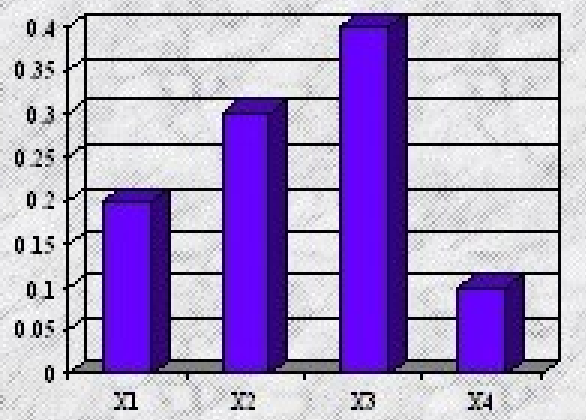
Variable Aleatoria Discreta

$$X \in \{x_1, x_2, x_3, \ldots, x_n\}$$

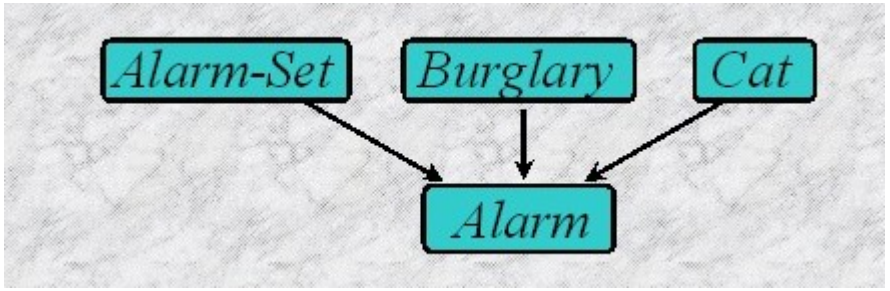
$$P(x_i) \geq 0$$

$$\sum_{i=1}^{n} P(x_i) = 1$$

X binary: $P(x) + P(\overline{x}) = 1$



Ejemplo: TPC de Variable Discreta

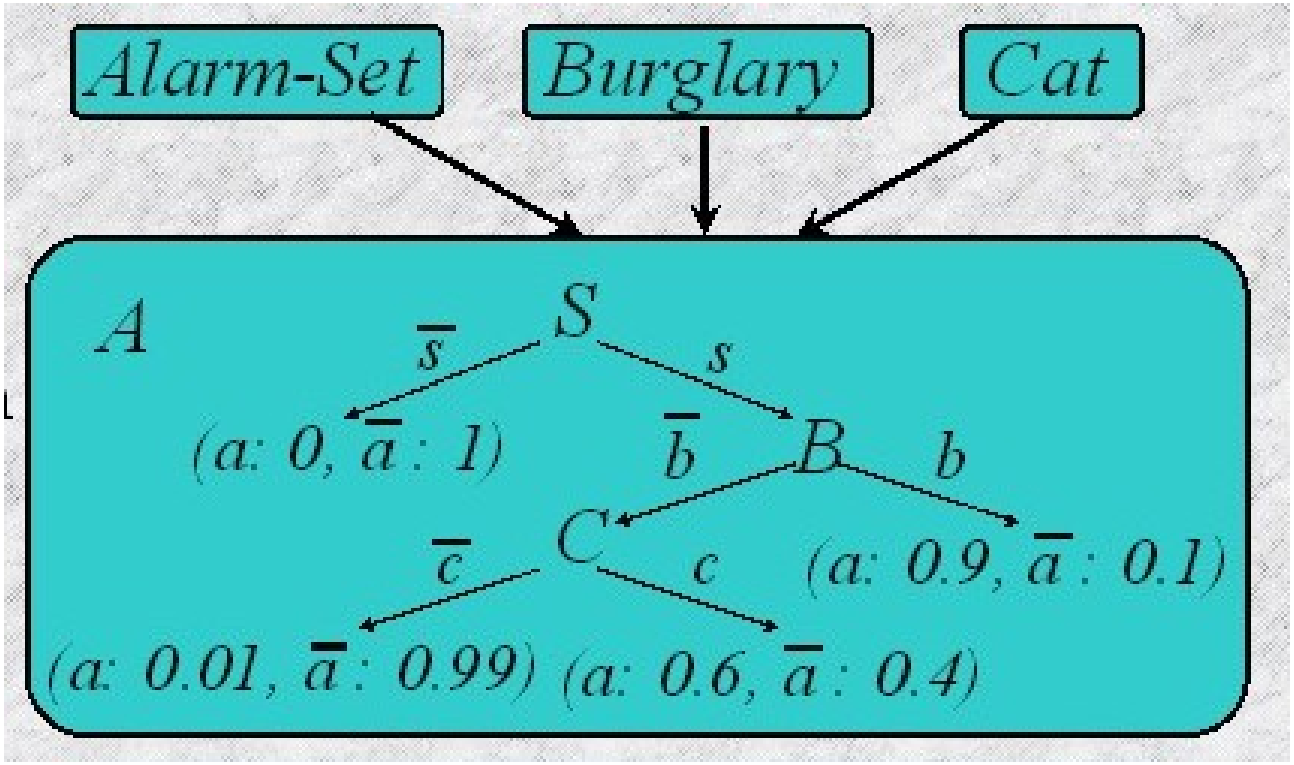


Dependencias del contexto:

- La alarma se puede apagar sólo si se ha activado
- Tanto B como C pueden activar la alarma
- Si B entra C se esconde y no activa la alarma

Ejemplo: TPC de Variable Discreta

Para el ejemplo, podemos codificar la función usando el Siguiente árbol:



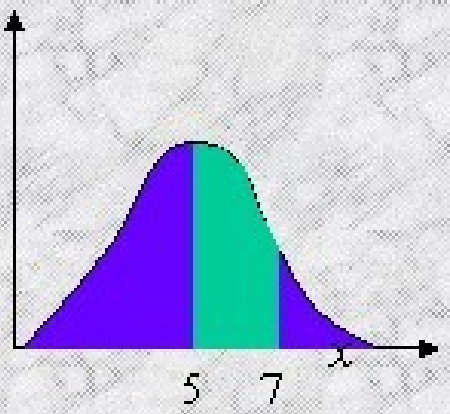
Variable Aleatoria Continua

Función de densidad

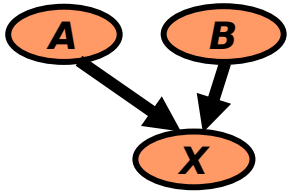
$$X \in [0,10] \qquad P(x) \geq 0$$

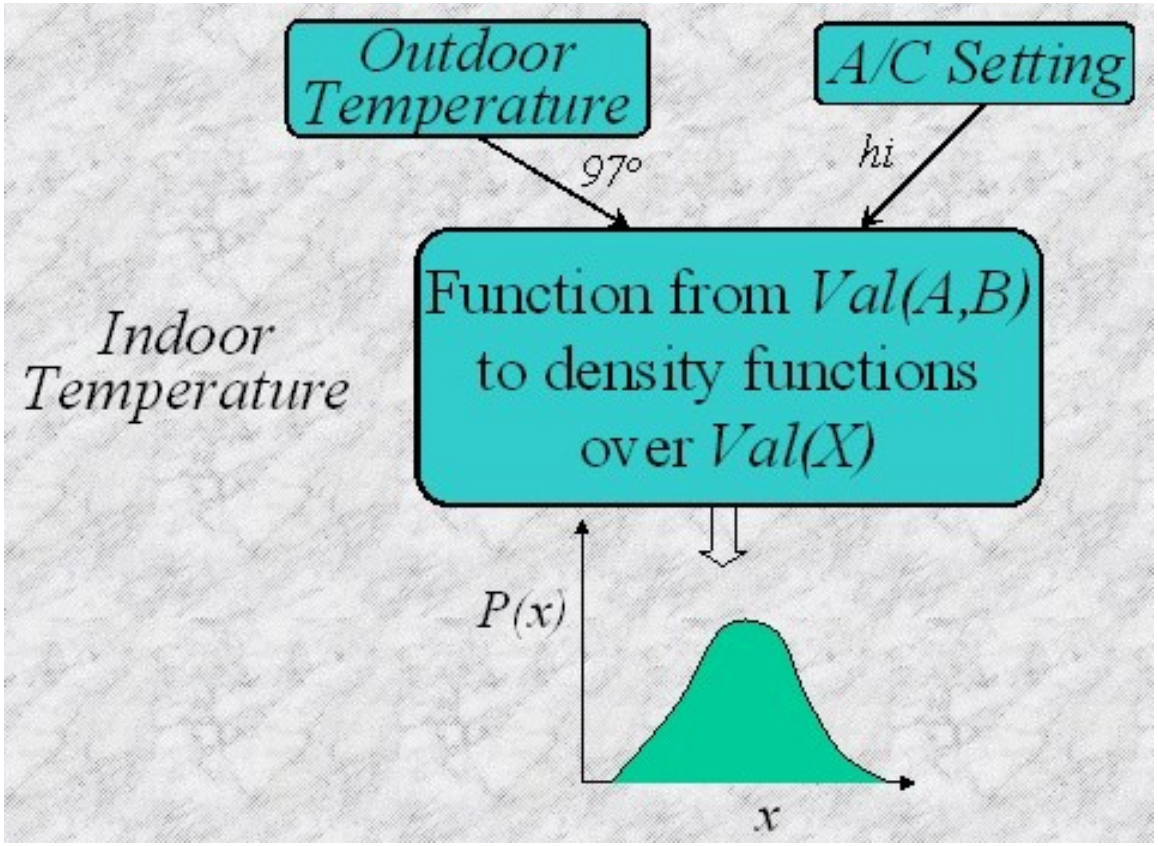
$$\int_0^{10} P(x)\,dx = 1$$

$$P(5 \leq x \leq 7) = \int_5^7 P(x)\,dx$$



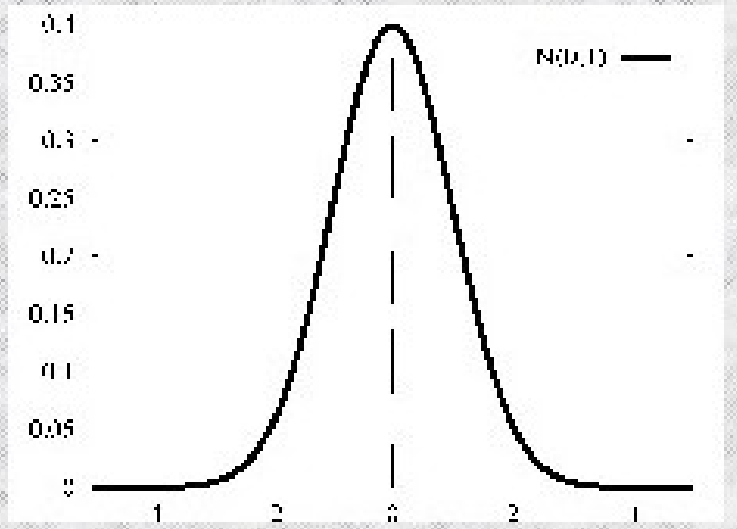
Ejemplo: TPC de Variable Continua





TPC de Variable Continua: Variable con distribución Normal

$$P(x) = \underbrace{\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma}\right)}_{N(\mu,\ \sigma)}$$



Estimación de TPCs

- Se reduce a “estimación de parametros” de distribuciones de prob.
- Para cada entrada de la TPC, estimar los parámetros de su distr. asociada:
  - Variable continua: en general se supone distribución Normal
  - Variable discreta: en general se supone distribución Multinomial
    - Caso particular: variable binaria, distribución Binomial

Estimación de Parámetros

- Estimación basada en Máxima Verosimilitud
  - Los datos determinan los parámetros
  - Problema: sobreajuste.
- Estimación Bayesiana
  - Idea: tenemos información a priori de los parámetros (prior)
  - Los datos nos aportan información adicional para ajustar los parámetros

Estimación Basada en Máxima Verosimilitud

- Datos para estimar: $D=\begin{bmatrix} X[1] \\ \vdots \\ X[M] \end{bmatrix}$

- Función de verosimilitud:

$$L(\theta:D)=P(D|\theta)=\prod_{m} P(X[m]|\theta)$$

- Estimador de máxima verosimilitud

$$\hat{\theta}=ArgMax_{\theta}\, L(\theta:D)$$

Estimador de Máx. Verosimilitud: Distr. Multinomial

$$\theta = \theta_1, \ldots, \theta_K$$

$$L(\theta:D) = P(D|\theta) = \prod_m P(x[m]|\theta)$$

$$L(\theta:D) = \prod_{i=1}^{K} \theta_i^{N_i}$$

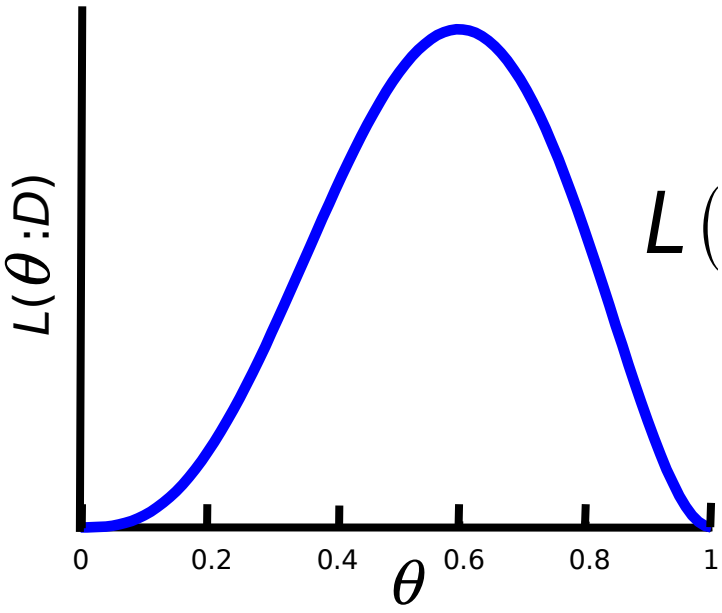
Conteo del i^{th} valor en D

Probabilidad del i^{h} valor

$$\hat{\theta} = \frac{N_1}{M}, \ldots, \frac{N_K}{M}$$

Ejemplo

- Variable binaria X : {h,t}
- Función de verosimilitud para: $D = \begin{matrix} h \\ t \\ t \\ h \\ h \end{matrix}$



$$L(\theta:D) = \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta$$

$$\theta = \frac{2}{5}, \frac{3}{5}$$

Estimador de Máx. Verosimilitud: Distribución Normal

$$\theta = \mu, \sigma^2$$

$$\hat{\theta} = \hat{\mu}, \hat{\sigma}^2$$

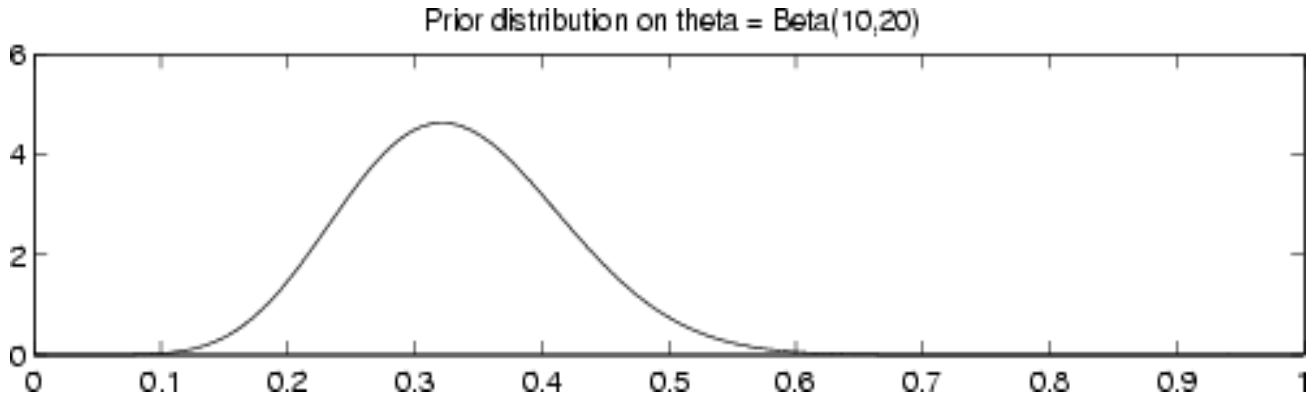
$$\hat{\mu} = \frac{1}{M} \sum_m x[m]$$

$$\hat{\sigma}^2 = \frac{1}{M} \sum_m \left(x[m] - \hat{\mu} \right)^2$$

Estimación Bayesiana (Ejemplo: Binomial)

- Información a priori:

$$P(\theta) \sim \text{Beta}(10,20) \qquad \frac{\theta^9(1-\theta)^{19}}{\int \theta^9(1-\theta)^{19}d\theta}$$



Esta distribución nos da la prob.
del parámetro dado que habíamos observamos 10 caras y 20 sellos a priori (por ejemplo en un experimento anterior o según un experto).

Estimación Bayesiana (Ejemplo: cont.)

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} = \frac{P(y|\theta)P(\theta)}{\int P(y|\theta)P(\theta)d\theta}$$

Estimación Bayesiana (cont.)

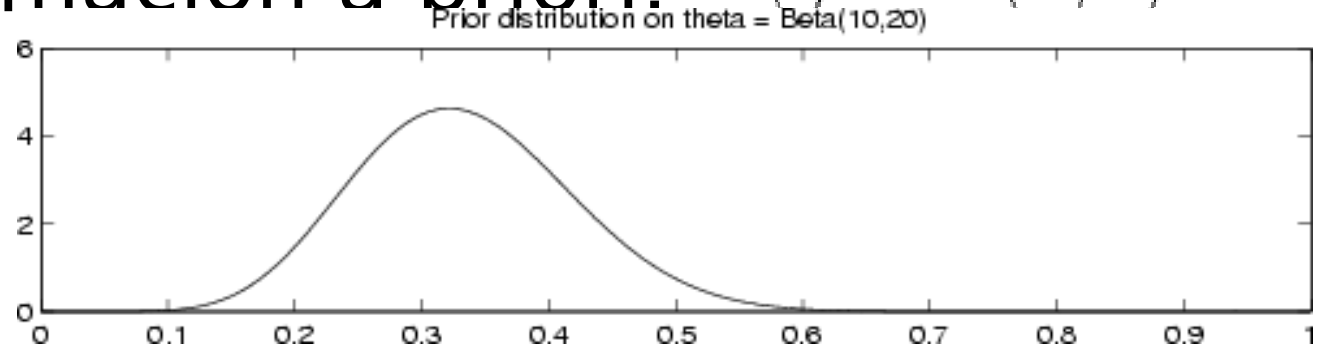
- Supongamos que observamos 50 caras y 50 sellos

$$P(\theta|y) = \frac{\theta^{59}(1-\theta)^{69}}{\int \theta^{59}(1-\theta)^{69} d\theta}$$

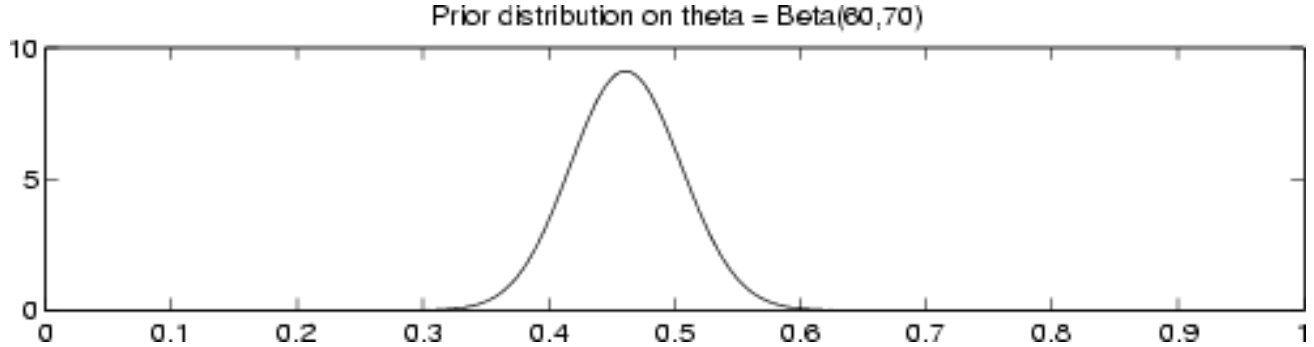
$$\text{Beta}(10+50, 20+50)$$

Estimación Bayesiana (Ejemplo: Binomial)

- Información a priori: $P(\theta) \sim \text{Beta}(10, 20)$



- Posterior $P(\theta|y) \quad \text{Beta}(10+50, 20+50)$

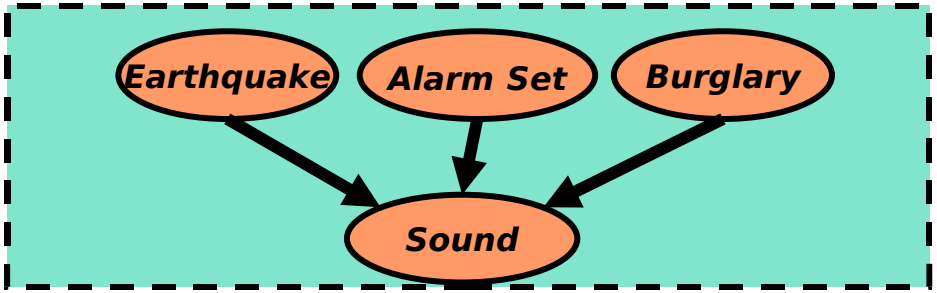


Construcción de una Red Bayesiana

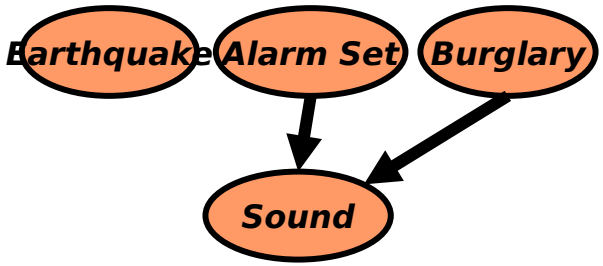
Veremos tres casos:

- Red y TPCs se definen por expertos
- Red predefinida y TPCs inducidas de los datos
- Red y TPCs inducidas de los datos

¿Por qué es importante la Estructura de la Red ?

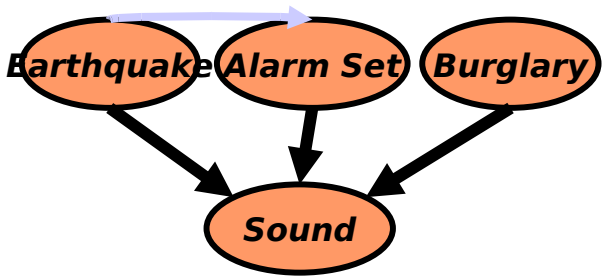


Arco faltante



Genera error que no se puede eliminar en la estimación de parámetros.

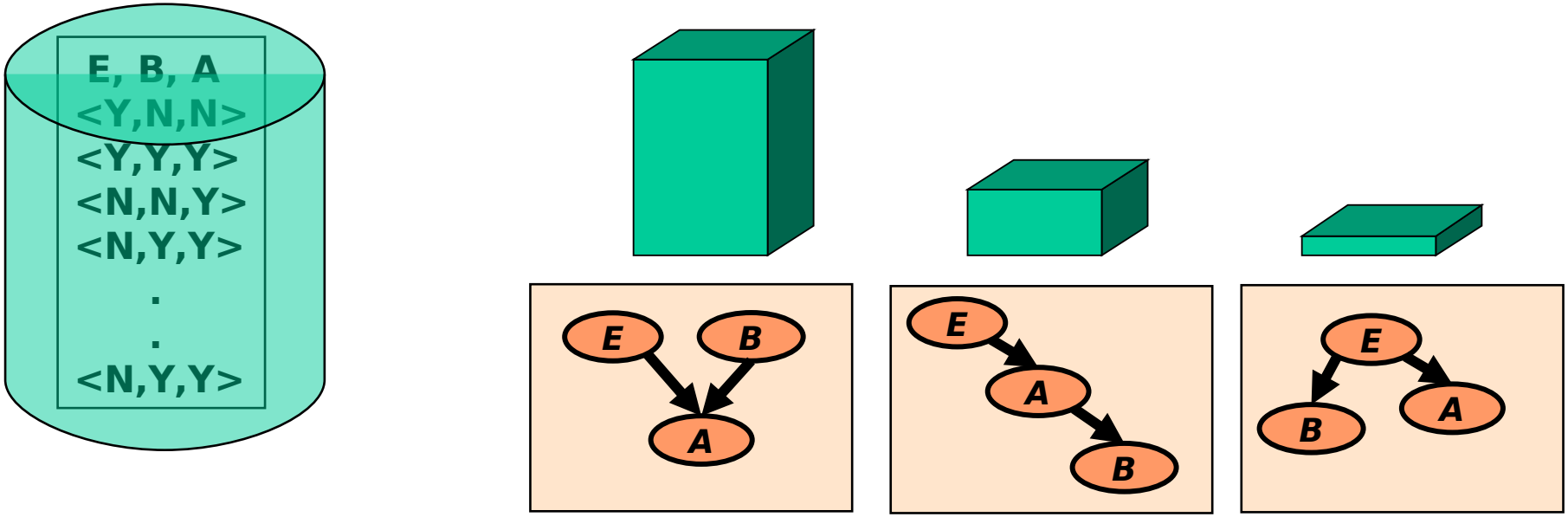
Arco innecesario



Incrementa el número de parámetros de la red

Inducción basado en Puntaje

Definir un puntaje (función) que indique que tan bien la red describe los datos



Buscar una red que maximice el puntaje

Espacio de Búsqueda

- Una red en el espacio de búsqueda se puede ver como un par:

$$B = (\mathbf{G}, \mathbf{P}_{G,D})$$

Grafo (estructura)

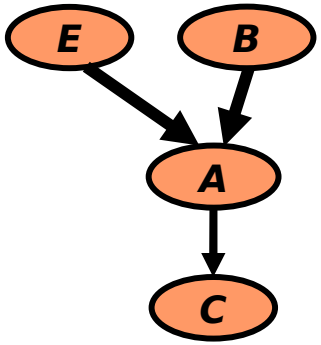
Probabilidad conjunta estimada en base a G y D

Espacio de Búsqueda

- Podemos suponer que dada una estructura y datos las probabilidades locales est'an determinadas.

Puntaje Basado en Verosimilitud: Ejemplo

- Red:



- Datos:

$$D=\begin{bmatrix} E[1] & B[1] & A[1] & C[1] \\ & & & \\ E[M] & B[M] & A[M] & C[M] \end{bmatrix}$$

$$L(G:D)=\prod_{m} P_{G,D}(E[m],B[m],A[m],C[m])$$

Puntaje Basado en Verosimilitud: Ejemplo

$$L(G:D)=\prod_m P_{G,D}(E[m],B[m],A[m],C[m])$$

$$\ell(\mathbf{G:D})=\log L(\mathbf{G:D})$$

$$\ell(G:D)=\sum_m \log P_{G,D}(E[m],B[m],A[m],C[m])$$

$$\ell(G:D)=M\sum_k P_D(E[k],B[k],A[k],C[k])\log P_{G,D}(E[k],B[k],A[k],C[k])$$

Probabilidad conjunta de los datos

Probabilidad conjunta de la red

Puntaje Basado en Verosimilitud

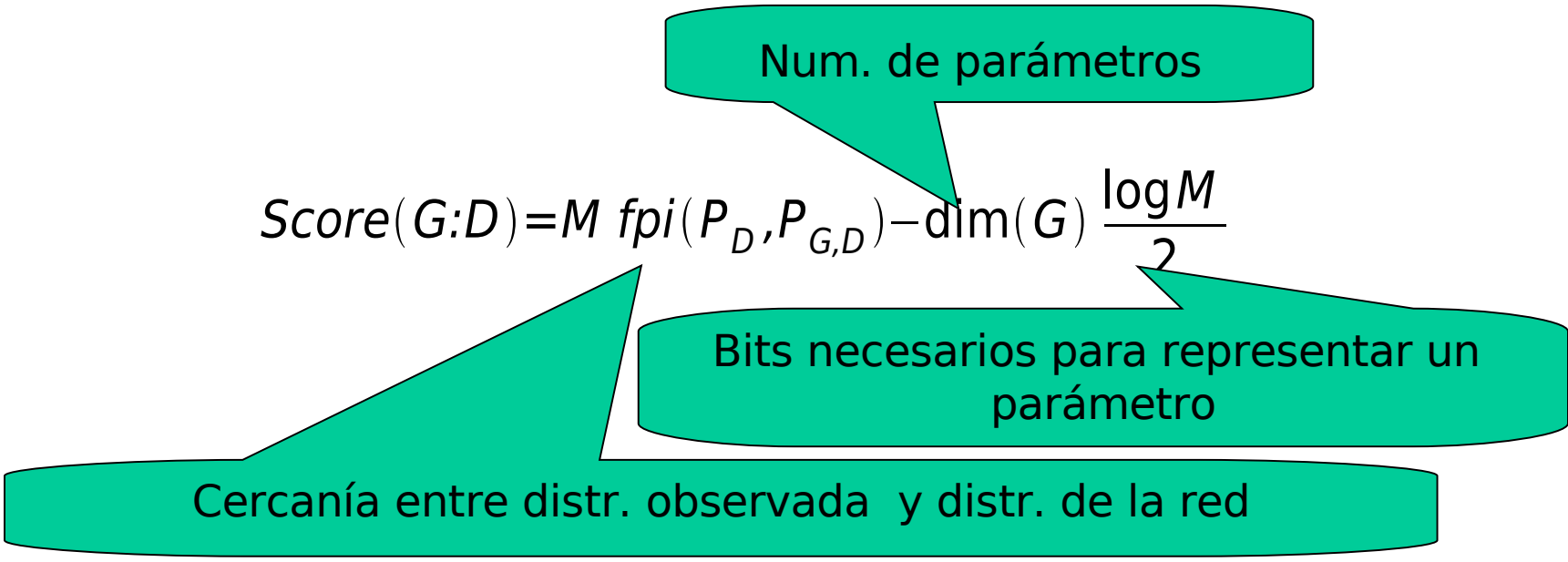
$$\ell(G:D) = M\, fpi(P_D, P_{G,D})$$

fpi: *función de pérdida de información.* Mide que tanto se asemeja la prob. conjunta de la red a la prob. conjunta de los datos.

Puntaje Basado en Verosimilitud

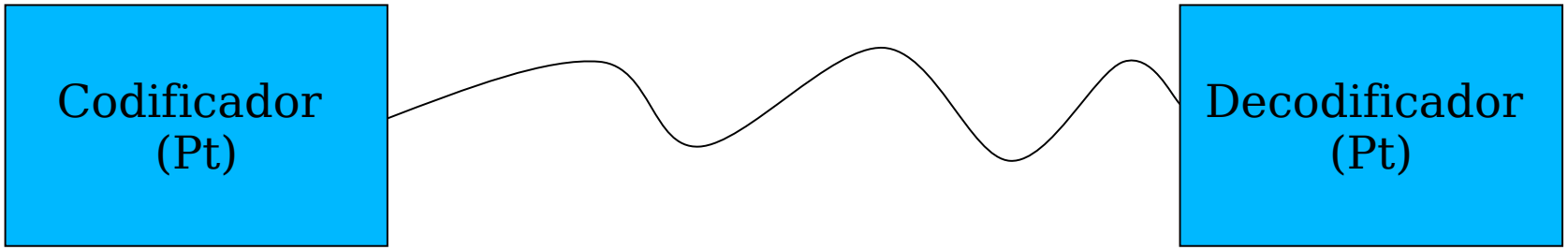
- Problema:
  - Es fácil demostrar que puntaje de verosimilitud obtenemos una red “completa”.
  - Problema de “sobreajuste”
- Soluciones:
  - Principio de Descripción Mínima:
    - Complementar fpi con penalización de complejidad de la red.

Puntaje Basado en Principio de Descripción Mínima



Este puntaje mide el costo de enviar los datos usando un código eficiente basado en la probabilidad conjunta de la red mas el costo de enviar la red.

Teoría de la Información: Arbol de un nodo (t)



X,C

a, c1

a, c1

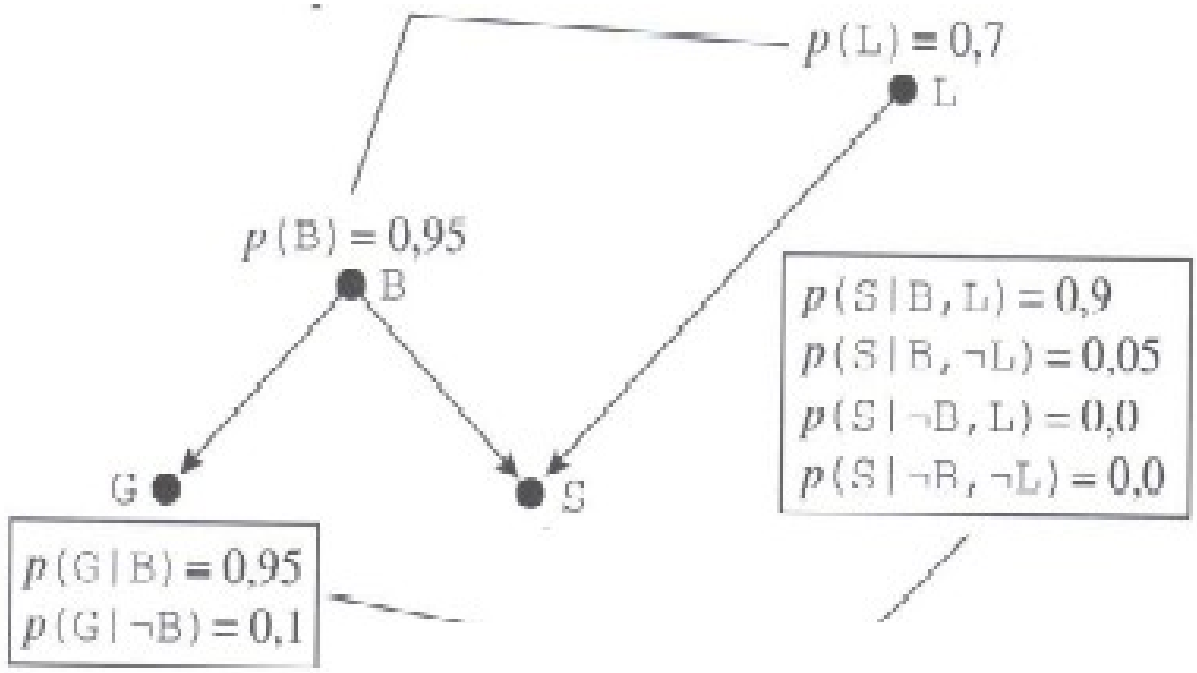
b, c2

b, c1

-Envío Red Bayesina (Pt)
-Envío los datos

X,C

Cantidad de bits a enviar = M fpi(PD,PG) + costo enviar Red



G	M	B	L	N.º de instancias
Verdadero	*Verdadero*	*Verdadero*	*Verdadero*	54
Verdadero	*Verdadero*	*Verdadero*	*Falso*	1
Verdadero	*Falso*	*Verdadero*	*Verdadero*	7
Verdadero	*Falso*	*Verdadero*	*Falso*	27
Falso	*Verdadero*	*Verdadero*	*Verdadero*	3
Falso	*Falso*	*Verdadero*	*Falso*	2
Falso	*Falso*	*Falso*	*Verdadero*	4
Falso	*Falso*	*Falso*	*Falso*	2
				100

$$Score(G{:}D) = 196{,}68 - 8\,\frac{\log 100}{2} = 223{,}26$$

Búsqueda de Estructura como Problema de Optimización

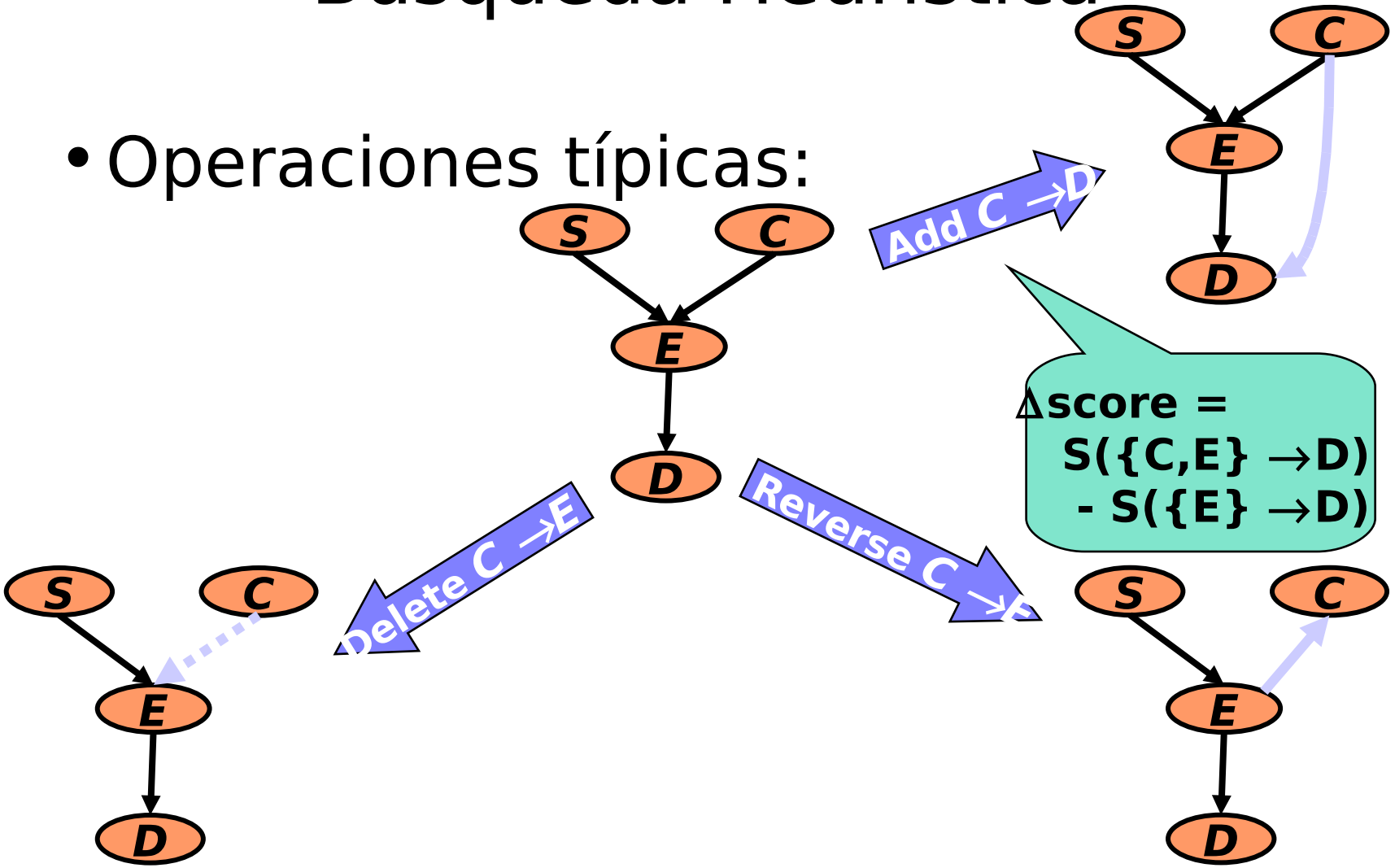
- Entrada:
  - Datos de Entrenamiento
  - Función de puntaje
  - Conjunto posible de estructuras
- Salida:
  - Red que maximiza el puntaje
- Búsqueda de estructuras con puntaje máximo, para redes con al menos $k > 1$ padres por nodo es NP-duro.
- Para k=1 (árboles) el problema se puede resolver en tiempo polinomial.
- Arboles tienen pocos parámetros por lo que en general evitan sobreajuste

Búsqueda para Estructuras Generales

- Se define un espacio de búsqueda
  - Estados son posibles estructuras
  - Operadores permiten moverse entre estados
- Navegar el espacio en busca de estructuras con puntaje alto:
  - Greedy hill-climbing
  - Tabu Search
  - Simulated Annealing
  - ...

Búsqueda Heurística

- Operaciones típicas:



Búsqueda de Estructura en Arboles

$$L(G:D)=\prod_m P(x_1[m],\ldots,x_n[m]:\Theta)$$

$$\prod_i \prod_m P(x_i[m]|Pa_i[m]:\Theta_i)$$

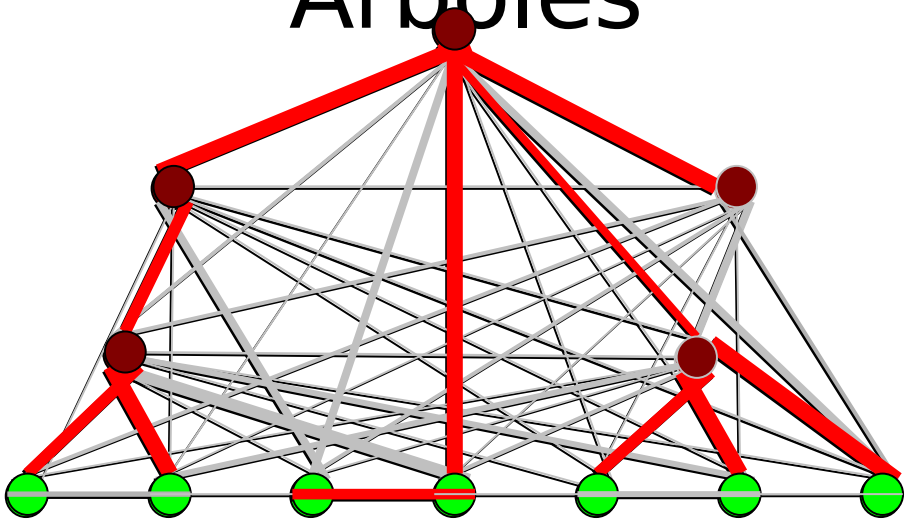
$$\prod_i L_i(edge_i:D)$$

$$Score(G:D)=\sum_i Score(edge_i:D)$$

$$Score(G:D)=\sum_i (Score(edge_i:D)-Score(X_i:D))-\sum_i Score(X_i:D)$$

Puntaje = suma de los ptjes de arcos + constante

Búsqueda de Estructura en Arboles



- Sea $w(j \rightarrow i) = Score(X_j \rightarrow X_i) - Score(X_i)$
- Buscar árbol o bosque de peso máximo
  - Algoritmo estándar para max spanning tree algorithm.
  - Costo: $O(n^2 \log n)$
- Teorema: Este procedimiento encuentra árbol con puntaje máximo