Ch 04. Parametric Model

Part 1 Hidden Markov Model

Markov Chain

- **State** $\omega_i, i = 1, 2, \ldots$
- The state at time t $\omega(t)$
- State sequence in discrete time of length T

$$\omega^T = \{\omega(1), \omega(2), ..., \omega(T)\}$$

For example: $\omega^6 = \{\omega_1, \omega_4, \omega_2, \omega_2, \omega_1, \omega_4\}$

- Transition Probability (Matrix)

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

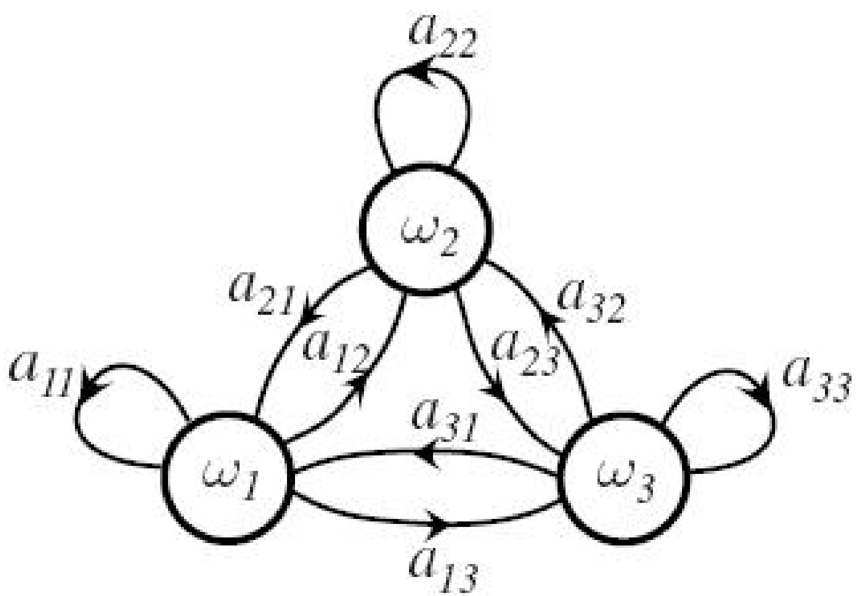
$$a_{ij} = P(\omega(t+1) = \omega_j \mid \omega(t) = \omega_i)$$

$$\sum_j a_{ij} = 1$$

a_{ij} is the transition probability from state ω_i to state ω_j

Markov Chain

- State Transition Diagram



Markov Chain

- j-order Markov process
  - The probability of being a certain state at next moment is only related to the nearest j states

$$P(\omega(t+1) \mid \omega(1),\omega(2),\ldots,\omega(t)) = P(\omega(t+1) \mid \underbrace{\omega(t-j+1),\omega(t-j+2),\ldots,\omega(t)}_{\text{only related to the nearest j states}})$$

- First-order Markov process
  - The probability of being a state at any moment is only related to the state at the previous moment

$$P(\omega(t+1) \mid \omega(1),\omega(2),\ldots,\omega(t)) = P(\omega(t+1) \mid \underbrace{\omega(t)}_{\substack{\text{only related to the state}\\ \text{at the previous moment}}})$$

Hidden Markov Model

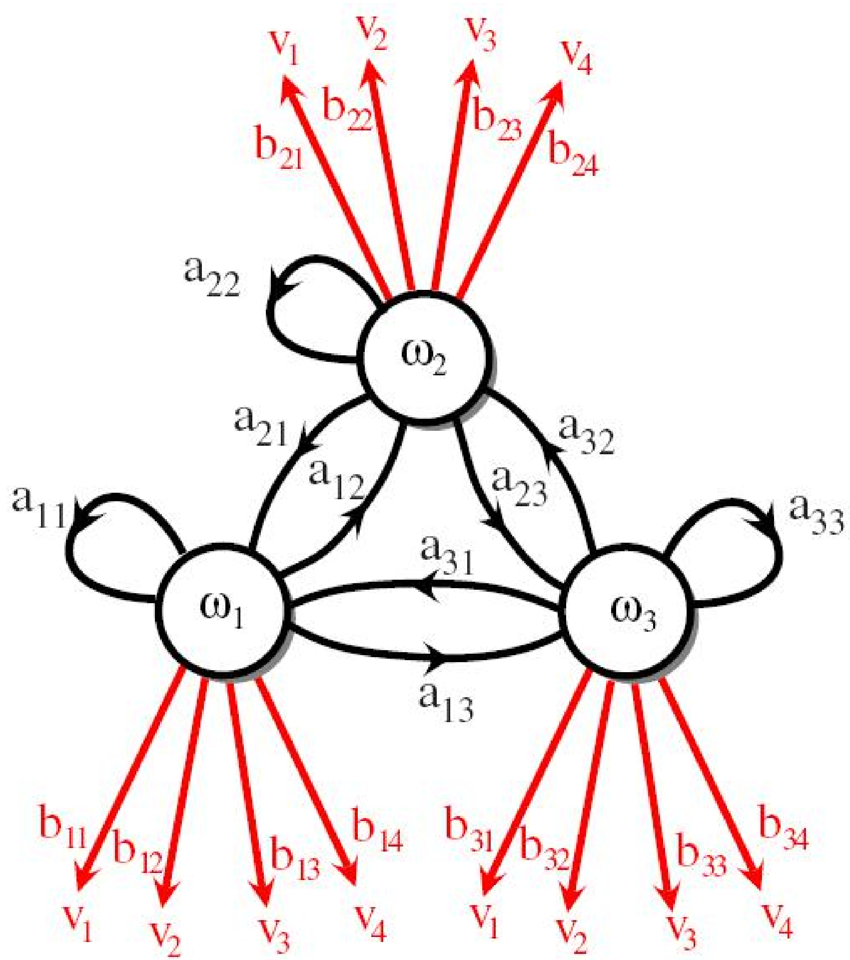
- **Hidden Markov Model** （ abbreviated to **HMM, 隐马尔可夫模型**）
- State is invisible
- At time t, the hidden state excites the visible symbol $x(t)$ with a certain probability, whose value is expressed as $v_1, v_2, v_3, \ldots$
- A sequence of visible symbols in discrete time of length T $\mathbf{X}^T = \{x(1), x(2), \ldots, x(T)\}$

  For example：$\mathbf{X}^6 = \{v_5, v_1, v_1, v_5, v_2, v_3\}$
- The probability of observing visible symbols

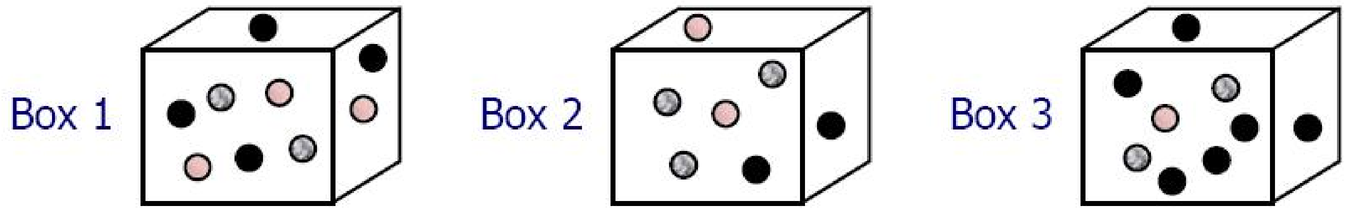
$$b_{jk} = P(x(t) = v_k \mid \omega(t) = \omega_j) \qquad \sum_k b_{jk} = 1$$

Hidden Markov Model

- State Transition Diagram



One example



- The box number is not visible
- Take one small ball out of any box at a time
- Hidden State: box number
- Visible symbol: small ball
- The probability of getting various small balls out of box i

- What's the probability of getting a particular sequence of small balls?

Symbolic Representation of Discrete HMM

- Hidden state set

  $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$

- Visible state set

  $V = \{v_1, v_2, \ldots, v_m\}$

- State sequence

  $\boldsymbol{\omega} = \omega(1)\ \omega(2) \cdots \omega(T)$

- Observed sequence

  $\mathbf{X} = x(1)\ x(2) \cdots x(T)$

- State transition probability

  $\mathbf{A} = \{a_{ij} \mid a_{ij} = P(\omega(t+1) = \omega_j \mid \omega(t) = \omega_i)\}$

- The probability of observing a visible sign

  $\mathbf{B} = \{b_{jk} \mid b_{jk} = P(x(t) = v_k \mid \omega(t) = \omega_j)\}$

- Initial state probability

  $\mathbf{\Pi} = \{\pi_i \mid \pi_i = P(\omega(1) = \omega_i)\}$

Complete HMM parameter vector

$\boldsymbol{\theta} = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$

Three Core Problems of HMM

- **Valuation problem**
  - **Given**
    - A specific symbol sequence **X** is observed
    - Parameter vector of HMM $\boldsymbol{\theta}$
  - **To solve**
    - Likelihood function $P(\mathbf{X} \mid \boldsymbol{\theta})$
- **Decoding problem**
  - **Given**
    - A specific symbol sequence **X** is observed
    - Parameter vector of HMM $\boldsymbol{\theta}$
  - **To solve**
    - The hidden state sequence most likely to produce **X**

Three Core Problems of HMM

- **Learning (or parameter estimation) problem**
  - **Given**
    - A specific symbol sequence **X** is observed
  - **To Solve**
    - The estimated value of the model parameter vector $\boldsymbol{\theta}$

      For example：ML estimation

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{X} \mid \boldsymbol{\theta})$$

Summary

- First-order Markov chain
- Hidden Markov Model（HMM）
- Three core problems of HMM
  - Valuation Problem
    - **HMM forward algorithm**
    - **HMM backward algorithm**

Valuation Problem

- Directly calculate the probability that the HMM model produces a symbol sequence X of visible length T

$$P(\mathbf{X} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{\omega}} P(\mathbf{X} \mid \boldsymbol{\omega}, \boldsymbol{\theta}) P(\boldsymbol{\omega} \mid \boldsymbol{\theta})$$

$$P(\mathbf{X} \mid \boldsymbol{\omega}, \boldsymbol{\theta}) = b_{\omega(1)x(1)} b_{\omega(2)x(2)} \cdots b_{\omega(T)x(T)}$$

$$P(\boldsymbol{\omega} \mid \boldsymbol{\theta}) = \pi_{\omega(1)} a_{\omega(1)\omega(2)} a_{\omega(2)\omega(3)} \cdots a_{\omega(T-1)\omega(T)}$$

$$P(\mathbf{X} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{\omega}} \prod_{t=1}^{T} a_{\omega(t-1)\omega(t)} b_{\omega(t)x(t)}$$

where $a_{\omega(0)\omega(1)}$ represents the initial probability $\pi_{\omega(1)}$ of the state $\omega(1)$

Suppose that there are c hidden states in HMM, the computational complexity is $O(c^T T)$!

For example：c=10，T=20，basic operations 10^{21} times!

Valuation Problem

- Solution
  - Recursive Computation

    The calculation at time T involves only the results of the previous step, $\omega(t)$, $\omega(t-1)$ and $x(t)$

    - **HMM forward algorithm**
    - **HMM backward algorithm**

Valuation Problem

- HMM forward algorithm

Define $\alpha_i(t)$:The probability that state is i at time t, and x(1), x(2), …… has been observed

- **Initialization**

for each hidden state i, calculate $\alpha_i(1) = \pi_i \, b_{ix(1)}$

- **Recursion**

for t=2 to T

For each hidden state j, calculate $\alpha_j(t) = \left[\sum_{i=1}^{c} \alpha_i(t-1)\, a_{ij} \right] b_{jx(t)}$

end

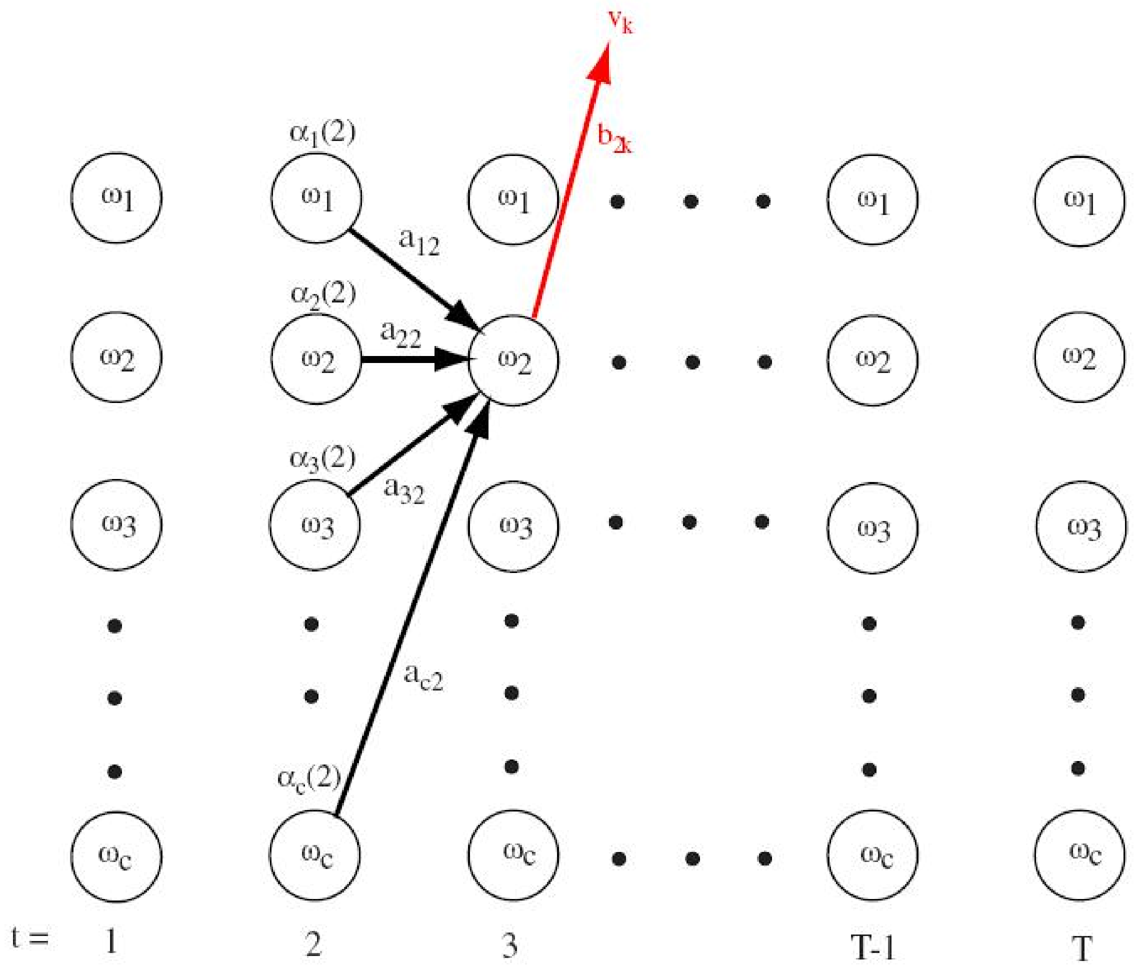
- **End**

$$P(\mathbf{X} \mid \boldsymbol{\theta}) = \sum_{i=1}^{c} \alpha_i(T)$$

Compute Complexity $O(c^2 T)$ $O(c^T T)$

Valuation Problem

- **HMM forward algorithm**



Valuation Problem

- HMM backward algorithm (time inversion version of forward algorithm)

  Define $\beta_i(t)$： The probability that state is i at time t, and x(T), x(T-1), ……x(t) has been observed in reverse

  - **Initialization**

    for each hidden state, calculate $\beta_i(T) = \dfrac{b_{ix(T)}}{c}$

    (suppose that the probability of each state at time T are the same)

  - **Recursion**

    for t=T-1 to 1

    For each hidden state i，compute $\beta_i(t) = \left[\sum_{j=1}^{c} a_{ij}\beta_j(t+1)\right] b_{ix(t)}$

    end

  - **End**

    $$P(\mathbf{X} \mid \boldsymbol{\theta}) = \sum_{i=1}^{c} \pi_i \beta_i(1)$$

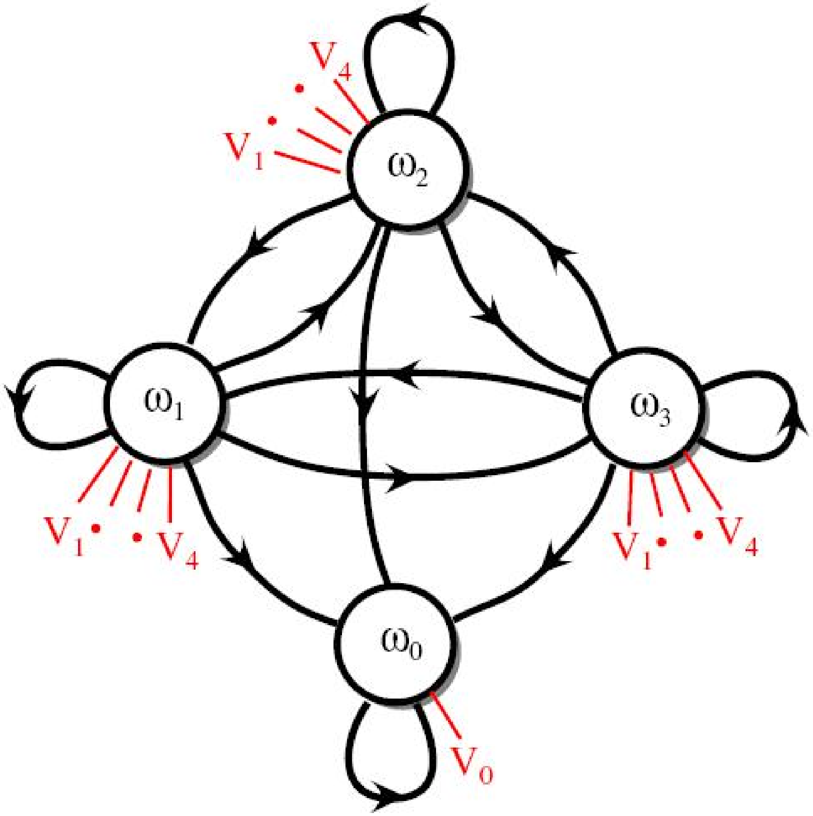
Compute Complexity $O(c^2T)$ $O(c^TT)$

Example

- HMM is

$$a_{ij} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.2 & 0.3 & 0.1 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{pmatrix}$$

$$b_{jk} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0.4 & 0.1 & 0.2 \\ 0 & 0.1 & 0.1 & 0.7 & 0.1 \\ 0 & 0.5 & 0.2 & 0.1 & 0.2 \end{pmatrix}$$



- ω_0: The absorbed state that the inevitable state at the end of the sequence. This state produces a unique particular visible symbol V_0, indicating the end of the HMM process

Example

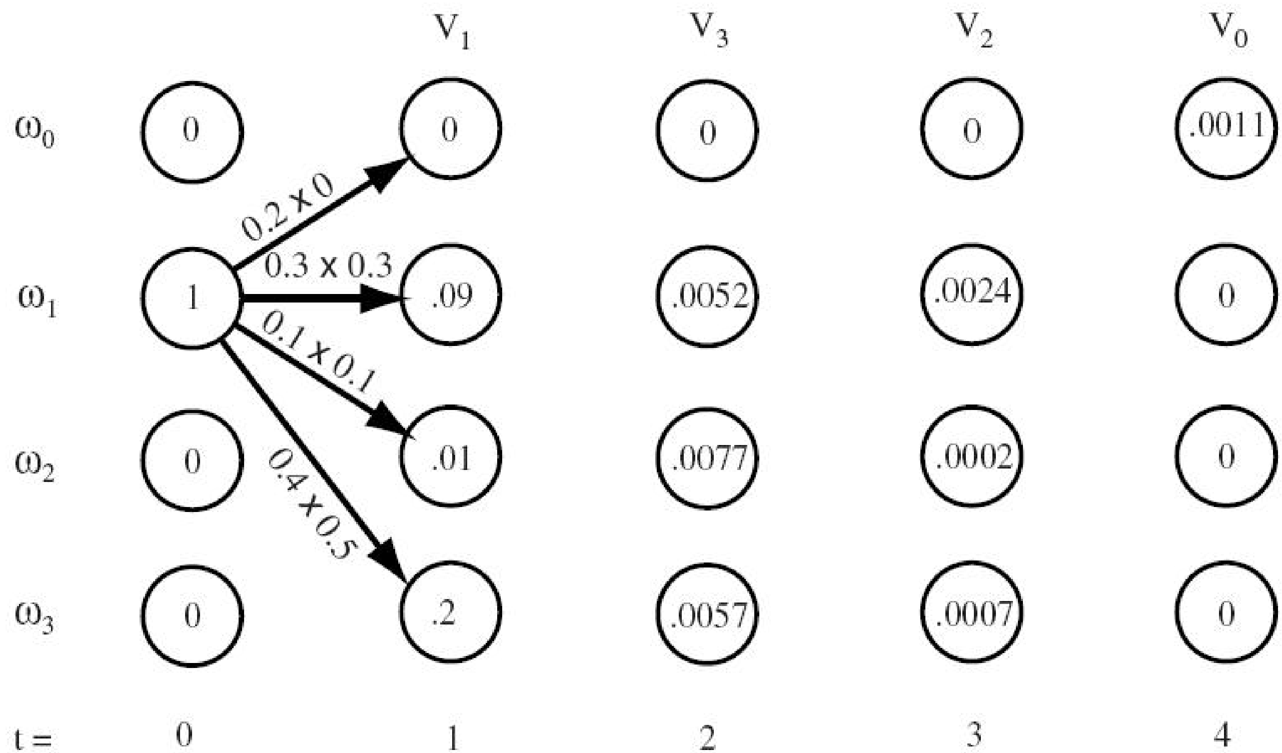
- Given state is ω_1 at time t，that is

$$\pi_0 = a_{10} = 0.2, \quad \pi_1 = a_{11} = 0.3,$$
$$\pi_2 = a_{12} = 0.1, \quad \pi_3 = a_{13} = 0.4$$

- The observed sequence is $\mathrm{V}^4 = \{v_1, v_3, v_2, v_0\}$
- Calculate the probability that the HMM produce this particular observation sequence

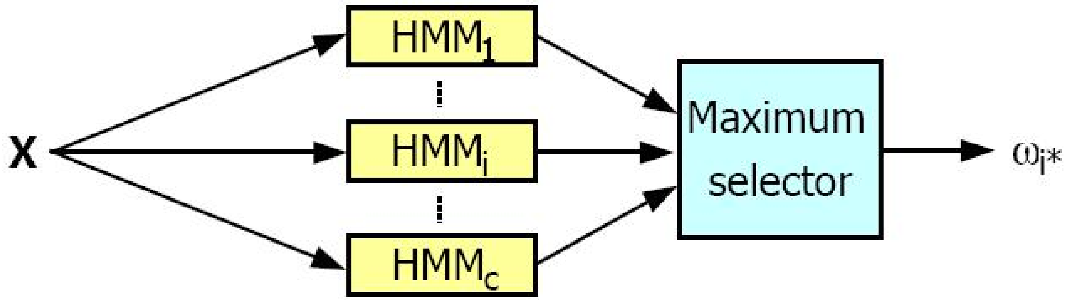
Example

- Solution



HMM for classification

- Build a HMM for each category

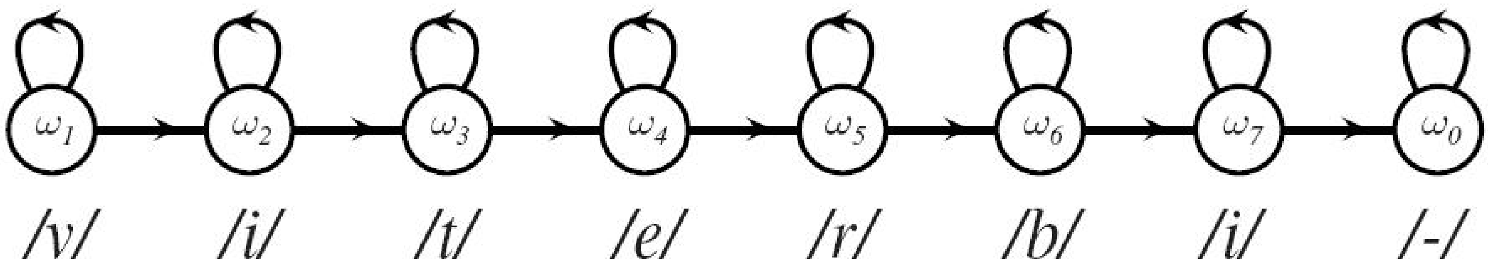


- Each HMM has its own parameter vector $\boldsymbol{\theta}_i$, which can be learned (estimated) from the samples belonging to category i
- Bayes Decision $P(\boldsymbol{\theta}_i \mid \mathbf{X}) = \dfrac{P(\mathbf{X} \mid \boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i)}{\sum_{i=1}^{c} P(\mathbf{X} \mid \boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i)}$
  - Decision Result

$$i^* = \arg\max_i (P(\mathbf{X} \mid \boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i))$$

HMM for Automatic Speech Recognition (ASR)

- left-to-right（从左到右）HMM



left-to-right HMM for pronunciation of viterbi

- Build a HMM for each word pronunciation，whose parameter is $\boldsymbol{\theta}_i$
- Use forward algorithm to calculate the class-conditional probability $P(\mathbf{X}|\boldsymbol{\theta}_i)$ of pronunciation sequence X
- $P(\boldsymbol{\theta}_i)$ depends on the language itself and the context semantics
- Use Bayes Formula to calculate the posterior probability $P(\boldsymbol{\theta}_i|\mathbf{X})$ of X
- The maximum posterior probability indicates the speech content

Decoding Problem

- Given a sequence of observations X^T, look for the most likely sequence of hidden states
- Method of exhaustion
  - Calculate the probabilities of all the possible sequences of hidden states
  - Compute complexity $O(c^T T)$

Decoding Problem

- **Viterbi algorithm**
  - **Initialization**

    for each hidden state i， compute $\delta_i(1) = \pi_i \, b_{ix(1)}$

  - **Recursion**

    for t=2 to T:

    For each hidden state j， compute

    $$\delta_j(t) = \left[\max_{1\le i\le c} \delta_i(t-1)\, a_{ij}\right] b_{jx(t)} \qquad \psi_j(t) = \arg\max_{1\le i\le c} \delta_i(t-1)\, a_{ij}$$

    end

  - **End**

    $$\omega^*(T) = \arg\max_{1\le i\le c} \delta_i(T)$$

    for t=T-1 to 1（path backtracking）:

    $$\omega^*(t) = \psi_{\omega^*(t+1)}(t+1)$$

    end

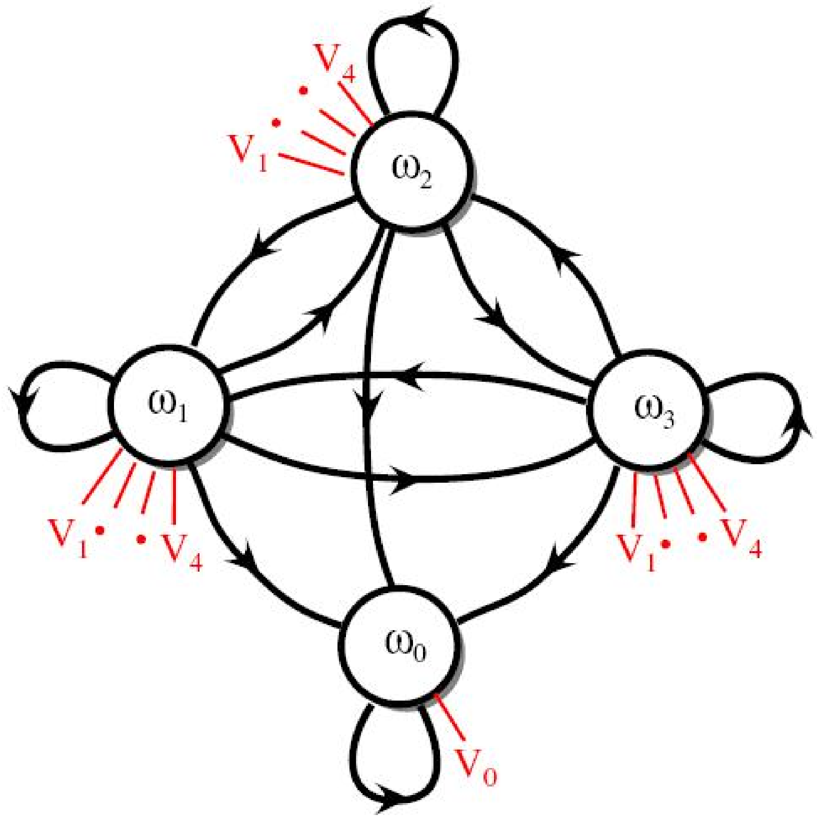
Compute Complexity $O(c^2T)$ $O(c^TT)$

Example

- HMM is

$$a_{ij} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.2 & 0.3 & 0.1 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{pmatrix}$$

$$b_{jk} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0.4 & 0.1 & 0.2 \\ 0 & 0.1 & 0.1 & 0.7 & 0.1 \\ 0 & 0.5 & 0.2 & 0.1 & 0.2 \end{pmatrix}$$



Example

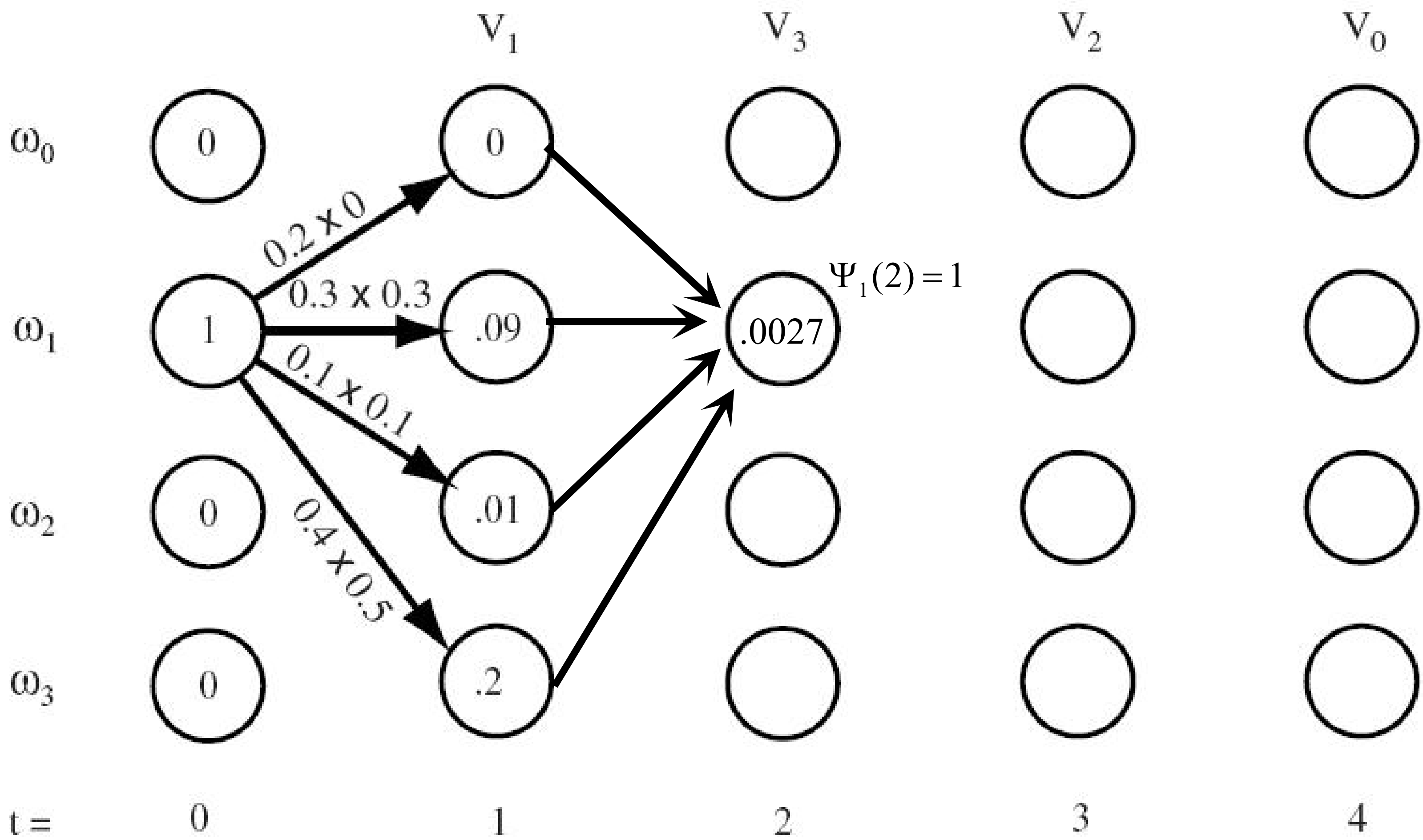
- Given that state is ω_1 at t=0,that is

$$\pi_0 = a_{10} = 0.2, \quad \pi_1 = a_{11} = 0.3,$$
$$\pi_2 = a_{12} = 0.1, \quad \pi_3 = a_{13} = 0.4$$

- The observed sequence is $\mathrm{V}^4 = \{v_1, v_3, v_2, v_0\}$
- Compute the most likely sequence of hidden states

Example

- Solution



Exercise： Fill out the diagram and trace back the optimal path

Decoding Problem

- For long sequences, the Viterbi algorithm may cause the computer underflow
- **Improvement:** Viterbi algorithm based on logarithm
  - Advantage
    - Change multiplication to addition
    - Avoid underflow
    - The results are same as Viterbi algorithm

$$\tilde{a}_{ij} = \log a_{ij}$$

$$\tilde{b}_{ix(t)} = \log b_{ix(t)}$$

$$\tilde{\pi}_i = \log \pi_i$$

$$\tilde{\delta}_i(t) = \log \delta_i(t)$$

Decoding Problem

- **Log-Viterbi algorithm**
  - **Initialization**

    for each hidden state i, compute $\tilde{\delta}_i(1) = \tilde{\pi}_i + \tilde{b}_{ix(1)}$

  - **Recursion**

    for t=2 to T:

    for each hidden state i, compute

    $$\tilde{\delta}_j(t) = \max_{1 \le i \le c} \left[\tilde{\delta}_i(t-1) + \tilde{a}_{ij} \right] + \tilde{b}_{jx(t)} \qquad \psi_j(t) = \arg\max_{1 \le i \le c} \left[\tilde{\delta}_i(t-1) + \tilde{a}_{ij} \right]$$

    end

  - **End**

    $$\omega^*(T) = \arg\max_{1 \le i \le c} \tilde{\delta}_i(T)$$

    for t=T-1 to 1（path backtracking）:

    $$\omega^*(t) = \psi_{\omega^*(t+1)}(t+1)$$

    end

Learning Problem

- Learning parameter vector θ of HMM from a set of training data $D=\{X_1, X_2, \ldots, X_n\}$
- There is no algorithm to determine the optimal HMM parameter based on the training set
- Common algorithms

forward-backward algorithm （**向前向后算法**）

also known as

Baum-Welch re-estimation algorithm

（ **Baum-Welch 重估计算法**）

- **Core idea**
  - Recursively update the HMM parameters to get the HMM parameters that best explain the training sample

Learning Problem

- **Baum-Welch re-estimation algorithm**
- Given X and $\boldsymbol{\theta}$, the posterior probability that is state i at time t and is state j at time t+1

$$\gamma_{ij}(t) = \frac{\alpha_i(t-1) a_{ij} b_{jk} \beta_j(t)}{P(\mathbf{X}^T \mid \boldsymbol{\theta})}$$

forward

backward

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_{ij}(t)}{\sum_{t=1}^{T} \sum_{k} \gamma_{ik}(t)}$$

$$\hat{b}_{jk} = \frac{\sum_{\substack{t=1 \\ v(t)=v_k}}^{T} \sum_{l} \gamma_{jl}(t)}{\sum_{t=1}^{T} \sum_{l} \gamma_{jl}(t)}$$

Learning Problem

- **forward-backward algorithm**
  - **Initiate** $\boldsymbol{\theta}$
  - **<u>repeat</u>**

    use **Baum-Welch re-estimation algorithm** to compute $\hat{\boldsymbol{\theta}}$

    based on $\boldsymbol{\theta}$ and **X**

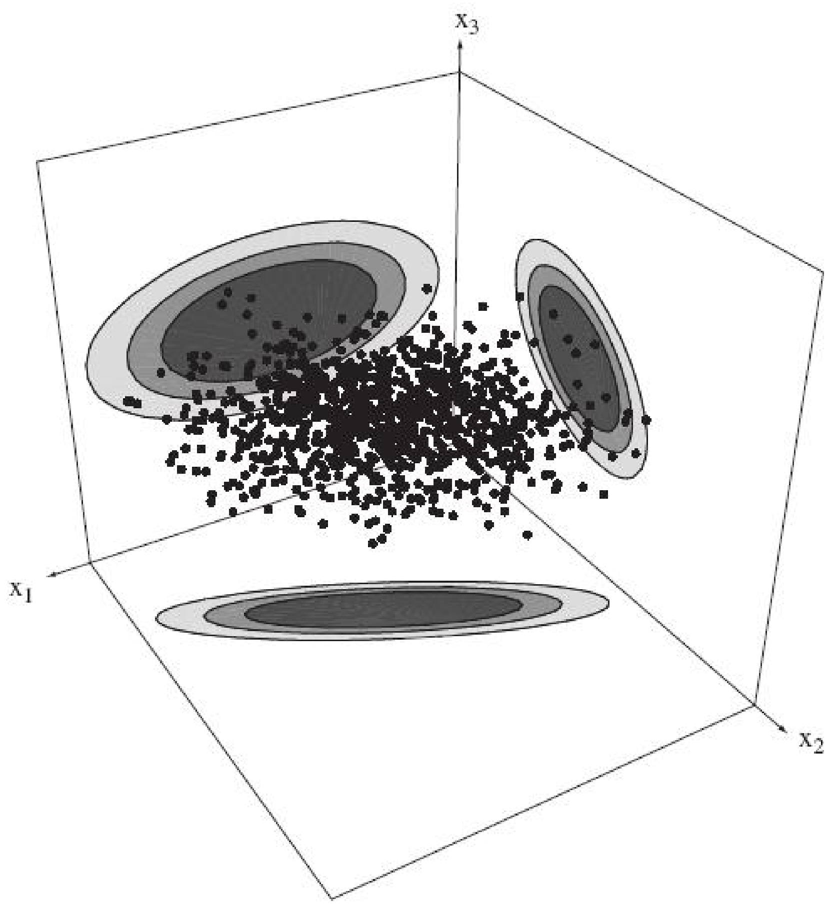
    $\boldsymbol{\theta} \leftarrow \hat{\boldsymbol{\theta}}$

    <u>until</u> $\boldsymbol{\theta}$ convergence
  - returns the result of parameter estimate $\boldsymbol{\theta}$

Part 2 Bayesian Belief Net

Feature correlation

- In some cases, the prior knowledge about the distribution is not directly in the form of probability distribution, but related to the statistical correlation (or independence) relationship between each characteristic component



X_1 and X_3 are statistically independent, but the other features are not

Example of Correlation

- state of the car
  - engine temperature
  - oil temperature
  - oil pressure
  - tire pressure
- correlation
  - oil pressure and tire pressure are independent of each other
  - oil temperature is related to engine temperature

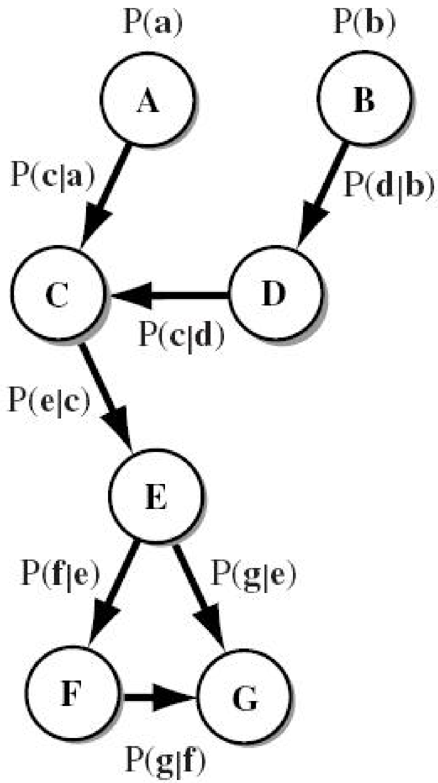
Bayesian Belief Net

- Represent the causal dependence between features by graph
  - **Bayesian belief net （贝叶斯置信网）**
  - **causal network （因果网）**
  - **belief net （置信网）**
- Direct Acyclic Graph (DAG)
  - The connection between nodes is directional
  - There is no cyclic path in the graph
- Only discrete cases are discussed

Bayesian Belief Net

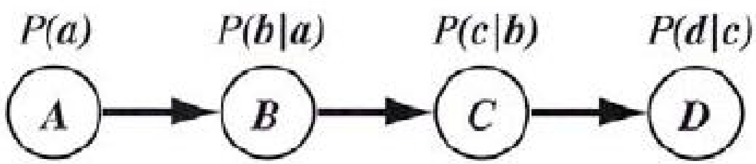
- Each node A, B, C... represents a system variable (feature)
  - The possible discrete values for each node
    - Values of A：$a_1, a_2, a_3, \ldots$
    - For example
      - A represents the state of the lamp
      - a_1=on， a_2=off，$P(a_1)=0.7$，$P(a_2)=0.3$
- Directed connections between nodes represent dependencies between variables
  - The connection from A to C represent

    $P(c_i \mid a_j)$ or $P(\mathbf{c} \mid \mathbf{a})$
- The state of any node can be inferred from the state of its adjacent nodes



Joint Probability

- Linear Chain



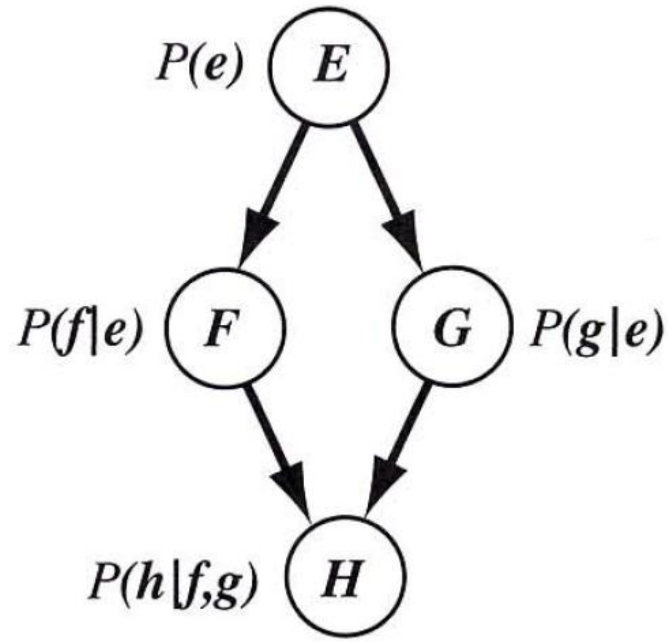
$$P(a,b,c,d) = P(a)P(b \mid a)P(c \mid b)P(d \mid c)$$

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

$$P(c,d) = P(d \mid c)\sum_a \sum_b P(a)P(b \mid a)P(c \mid b)$$

Joint Probability

- Simple circuit



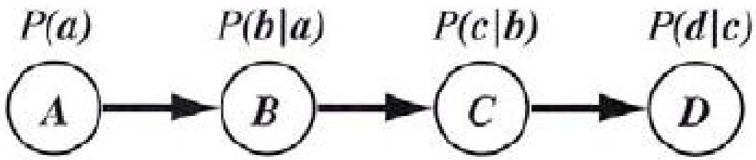
$$P(e,f,g,h) = P(e)P(f\,|\,e)P(g\,|\,e)P(h\,|\,f,g)$$

$$P(f,g,h) = P(h\,|\,f,g)\sum_{e} P(e)P(f\,|\,e)P(g\,|\,e)$$

$$P(g,h) = \sum_{e}\sum_{f} P(e)P(f\,|\,e)P(g\,|\,e)P(h\,|\,f,g)$$

The probability of any node taking a Specific value

- Linear chain



$$P(\mathbf{d}) = \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$$

$$= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a}) P(\mathbf{b}|\mathbf{a}) P(\mathbf{c}|\mathbf{b}) P(\mathbf{d}|\mathbf{c})$$

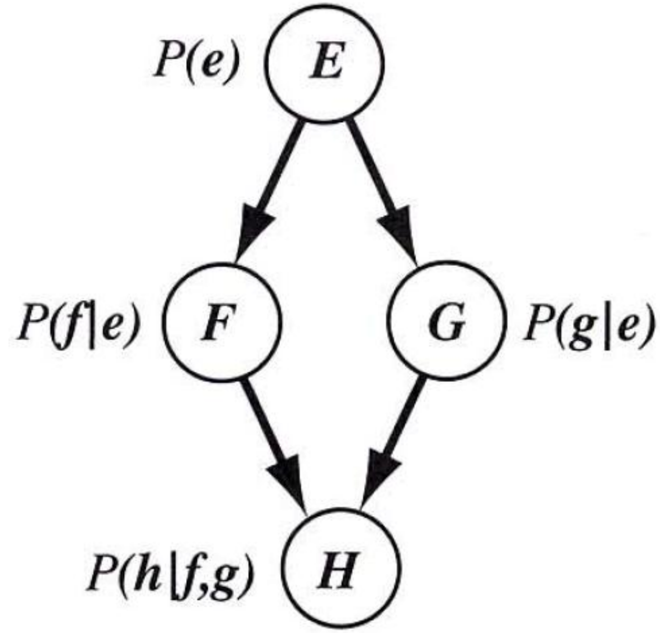
$$= \underbrace{\sum_{\mathbf{c}} P(\mathbf{d}|\mathbf{c}) \underbrace{\sum_{\mathbf{b}} P(\mathbf{c}|\mathbf{b}) \underbrace{\sum_{\mathbf{a}} P(\mathbf{b}|\mathbf{a}) P(\mathbf{a})}_{P(\mathbf{b})}}_{P(\mathbf{c})}}_{P(\mathbf{d})}$$

The probability of any node taking a Specific value

- Simple circuit

$$P(h) = \sum_{e,f,g} P(e,f,g,h)$$

$$= \sum_{e,f,g} P(e)P(f \mid e)P(g \mid e)P(h \mid f,g)$$



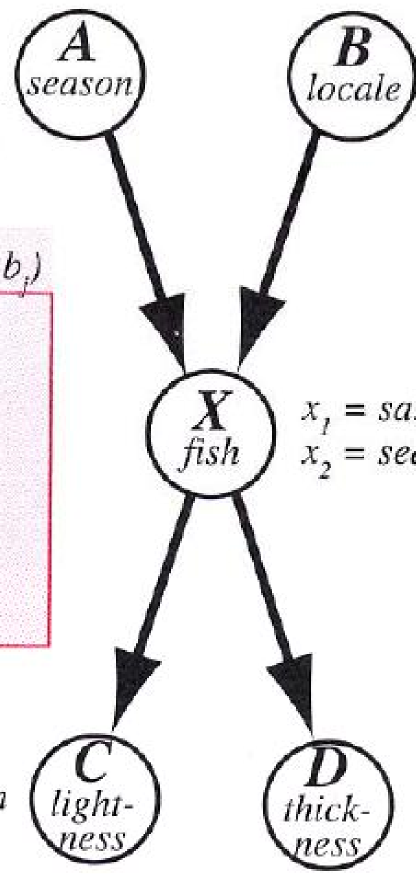
Example One

- fish classification belief net

$P(a)$

$P(a_1)$	$P(a_2)$	$P(a_3)$	$P(a_4)$
0.25	0.25	0.25	0.25

a_1 = winter
a_2 = spring
a_3 = summer
a_4 = autumn



b_1 = north Atlantic
b_2 = south Atlantic

$P(b)$

$P(b_1)$	$P(b_2)$
0.6	0.4

$P(x|a,b)$

	$P(x_1|a_i, b_j)$	$P(x_2|a_i, b_j)$
a_1, b_1	0.5	0.5
a_1, b_2	0.7	0.3
a_2, b_1	0.6	0.4
a_2, b_2	0.8	0.2
a_3, b_1	0.4	0.6
a_3, b_2	0.1	0.9
a_4, b_1	0.2	0.8
a_4, b_2	0.3	0.7

x_1 = salmon
x_2 = sea bass

$P(c|x)$

	$P(c_1|x_k)$	$P(c_2|x_k)$	$P(c_3|x_k)$
x_1	0.6	0.2	0.2
x_2	0.2	0.3	0.5

c_1 = light
c_2 = medium
c_3 = dark

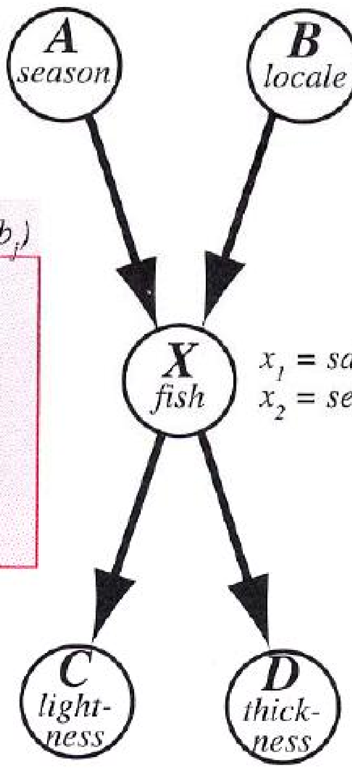
d_1 = wide
d_2 = thin

$P(d|x)$

	$P(d_1|x_k)$	$P(d_2|x_k)$
x_1	0.3	0.7
x_2	0.6	0.4

Example One

- Calculate the probability that "a fish caught in the North Atlantic in summer is a bass with a dim gloss and a narrow width"



a_1 = winter
a_2 = spring
a_3 = summer
a_4 = autumn

b_1 = north Atlantic
b_2 = south Atlantic

x_1 = salmon
x_2 = sea bass

c_1 = light
c_2 = medium
c_3 = dark

d_1 = wide
d_2 = thin

$P(a)$

$P(a_1)$	$P(a_2)$	$P(a_3)$	$P(a_4)$
0.25	0.25	0.25	0.25

$P(b)$

$P(b_1)$	$P(b_2)$
0.6	0.4

$P(x|a,b)$

	$P(x_1\|a_i, b_j)$	$P(x_2\|a_i, b_j)$
a_1, b_1	0.5	0.5
a_1, b_2	0.7	0.3
a_2, b_1	0.6	0.4
a_2, b_2	0.8	0.2
a_3, b_1	0.4	0.6
a_3, b_2	0.1	0.9
a_4, b_1	0.2	0.8
a_4, b_2	0.3	0.7

$P(c|x)$

	$P(c_1\|x_i)$	$P(c_2\|x_i)$	$P(c_3\|x_i)$
x_1	0.6	0.2	0.2
x_2	0.2	0.3	0.5

$P(d|x)$

	$P(d_1\|x_k)$	$P(d_2\|x_k)$
x_1	0.3	0.7
x_2	0.6	0.4

Example One

- Calculate the probability that "a fish caught in the North Atlantic in summer is a bass with a dim gloss and a narrow width"
  - Summer： a_3
  - North Atlantic ： b_1
  - dim gloss ： c_3
  - narrow width ： d_2
  - Bass： x_2

$$P(a_3,b_1,x_2,c_3,d_2) = P(a_3)P(b_1)P(x_2 \mid a_3,b_1)P(c_3 \mid x_2)P(d_2 \mid x_2)$$

$$= 0.25 \times 0.6 \times 0.6 \times 0.5 \times 0.4$$

$$= 0.018$$

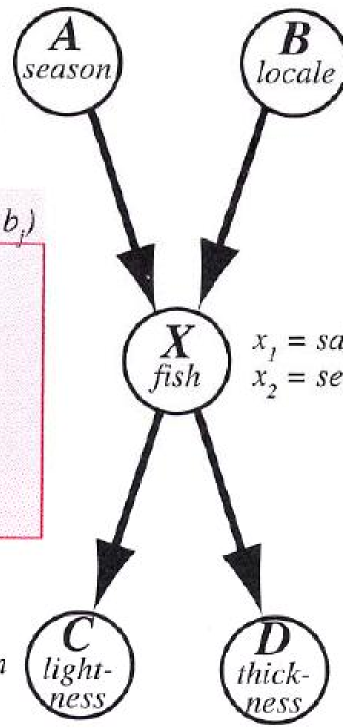
Example One

1. The probability of catching salmon in the South Atlantic in winter
2. The probability of catching bright bass in the South Atlantic
3. The probability of catching a wide, bright fish in the North Atlantic in summer

$P(a)$

$P(a_1)$	$P(a_2)$	$P(a_3)$	$P(a_4)$
0.25	0.25	0.25	0.25

a_1 = winter
a_2 = spring
a_3 = summer
a_4 = autumn



b_1 = north Atlantic
b_2 = south Atlantic

$P(b)$

$P(b_1)$	$P(b_2)$
0.6	0.4

$P(x|a,b)$

	$P(x_1|a_i, b_j)$	$P(x_2|a_i, b_j)$
a_1, b_1	0.5	0.5
a_1, b_2	0.7	0.3
a_2, b_1	0.6	0.4
a_2, b_2	0.8	0.2
a_3, b_1	0.4	0.6
a_3, b_2	0.1	0.9
a_4, b_1	0.2	0.8
a_4, b_2	0.3	0.7

x_1 = salmon
x_2 = sea bass

$P(c|x)$

	$P(c_1|x_k)$	$P(c_2|x_k)$	$P(c_3|x_k)$
x_1	0.6	0.2	0.2
x_2	0.2	0.3	0.5

c_1 = light
c_2 = medium
c_3 = dark

d_1 = wide
d_2 = thin

$P(d|x)$

	$P(d_1|x_k)$	$P(d_2|x_k)$
x_1	0.3	0.7
x_2	0.6	0.4

Evidence

- Given the values of variables other than the target variable X, determine the probability of the other variables
- Evidence $\{e_{\mathbf{A}}, e_{\mathbf{B}}, e_{\mathbf{C}}, e_{\mathbf{D}}\}$， where e_i indicate the value of the variable
- For example， fish classification belief net
  - Available evidence $\mathbf{e} = \{e_A, e_B, e_C, \ldots\}$
    - $e_{\mathbf{A}}$： winter now

      $P(a_1|e_{\mathbf{A}}) = 1 \quad P(a_i|e_{\mathbf{A}}) = 0 \text{ for } i = 2, 3, 4$
    - $e_{\mathbf{B}}$： fishermen prefer the South Atlantic

      $P(b_1|e_{\mathbf{B}}) = 0.2 \qquad P(b_2|e_{\mathbf{B}}) = 0.8$
    - $e_{\mathbf{C}}$： the fish has a lighter sheen

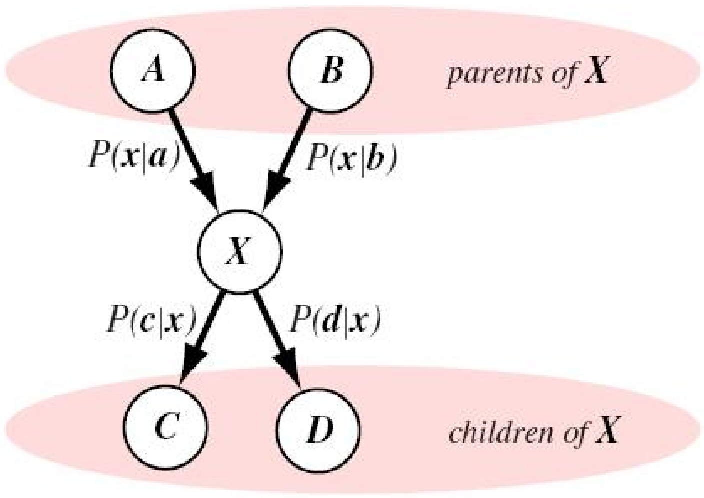
      $P(e_{\mathbf{C}}|c_1) = 1 \qquad P(e_{\mathbf{C}}|c_2) = 0.5 \qquad P(e_{\mathbf{C}}|c_3) = 0$
    - $e_{\mathbf{D}}$： the width cannot be measured because of occlusion

      $P(e_{\mathbf{D}}|d_1) = P(e_{\mathbf{D}}|d_2)$

Pay attention to the position of e_i !

Confidence

- Consider a certain node X
- The set of nodes before X is called the parent node P of X and the set of nodes after X is called the child node C of X
- For example：
  - parent node of X：{A，B}
  - child node of X ：{C，D}



- When estimating the probability of X, the parent node and the child node of X should be treated differently
  - Evidence **e**：Values of variables at nodes other than X
  - Given e, the Confidence Belief of x = (x1, x2...)

$$P(\mathbf{x}|\mathbf{e}) \propto P(\mathbf{e}^{\mathcal{C}}|\mathbf{x})P(\mathbf{x}|\mathbf{e}^{\mathcal{P}})$$

  - Must be normalized so that the sum of the probabilities of all values of x is 1

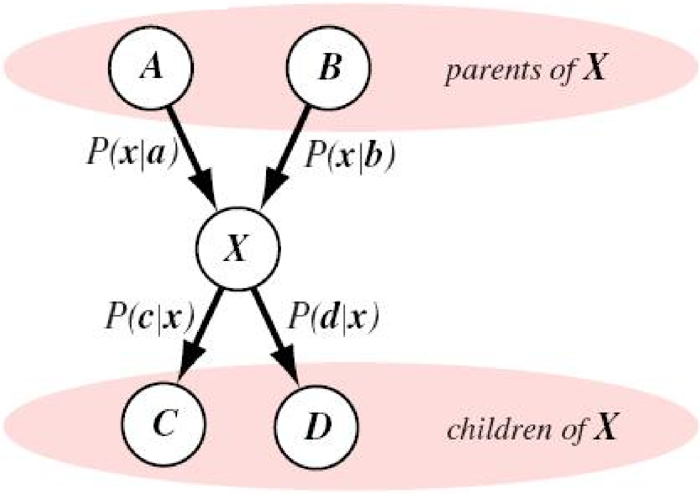
Confidence

- For child node of X
  - Suppose there are no connections between child nodes

$$
\begin{aligned}
P(\mathbf{e}^{\mathcal{C}}|\mathbf{x}) &= P(\mathbf{e}_{\mathcal{C}_1}, \mathbf{e}_{\mathcal{C}_2}, ..., \mathbf{e}_{\mathcal{C}_{|\mathcal{C}|}}|\mathbf{x}) \\
&= P(\mathbf{e}_{\mathcal{C}_1}|\mathbf{x})P(\mathbf{e}_{\mathcal{C}_2}|\mathbf{x}) \cdots P(\mathbf{e}_{\mathcal{C}_{|\mathcal{C}|}}|\mathbf{x}) \\
&= \prod_{j=1}^{|\mathcal{C}|} P(\mathbf{e}_{\mathcal{C}_j}|\mathbf{x}),
\end{aligned}
$$

  - For exmple：

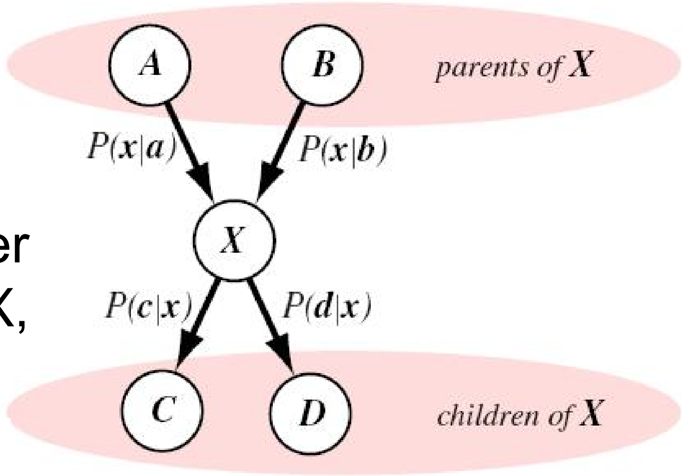
$$P(\mathbf{e}_{\mathbf{C}}, \mathbf{e}_{\mathbf{D}}|\mathbf{x}) = P(\mathbf{e}_{\mathbf{C}}|\mathbf{x})P(\mathbf{e}_{\mathbf{D}}|\mathbf{x})$$



Confidence

- For parent node of X
  - Suppose there are no connections between parent nodes

$$
\begin{aligned}
P(\mathrm{x}|\mathrm{e}^{\mathcal{P}}) &= P(\mathrm{x}|\mathrm{e}_{\mathcal{P}_1}, \mathrm{e}_{\mathcal{P}_2}, ..., \mathrm{e}_{\mathcal{P}_{|\mathcal{P}|}}) \\
&= \sum_{all\ \ i,j,...,k} P(\mathrm{x}|\mathcal{P}_{1i}, \mathcal{P}_{2j}, ..., \mathcal{P}_{|\mathcal{P}|k}) P(\mathcal{P}_{1i}, \mathcal{P}_{2j}, ..., \mathcal{P}_{|\mathcal{P}|k}|\mathrm{e}_{\mathcal{P}_1}, ..., \mathrm{e}_{\mathcal{P}_{|\mathcal{P}|}}) \\
&= \sum_{all\ \ i,j,...,k} P(\mathrm{x}|\mathcal{P}_{1i}, \mathcal{P}_{2j}, ..., \mathcal{P}_{|\mathcal{P}|k}) P(\mathcal{P}_{1i}|\mathrm{e}_{\mathcal{P}_1}) \cdots P(\mathcal{P}_{|\mathcal{P}|k}|\mathrm{e}_{\mathcal{P}_{|\mathcal{P}|k}}),
\end{aligned}
$$



- $\mathcal{P}_{mn}$ represents the value of the parent node $\mathcal{P}_m$ in state n
- Ignore node interdependencies other than the parent and child nodes of X, and simplify above equation

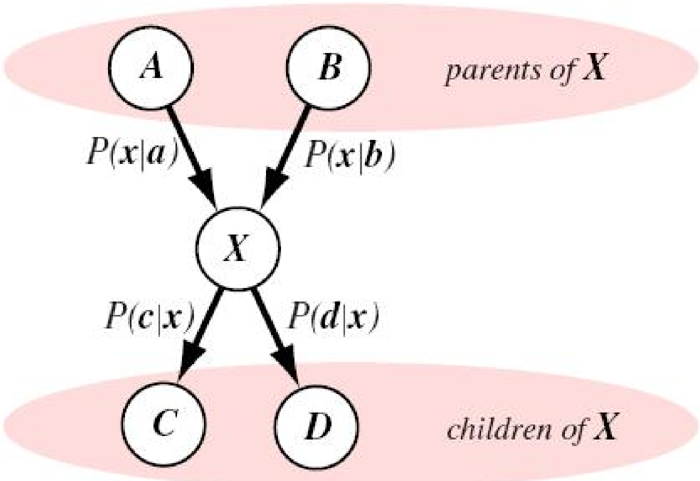
$$
P(\mathrm{x}|\mathrm{e}^{\mathcal{P}}) = \sum_{all\ \ \mathcal{P}_{mn}} P(\mathrm{x}|\mathcal{P}_{mn}) \prod_{i=1}^{|\mathcal{P}|} P(\mathcal{P}_i|\mathrm{e}_{\mathcal{P}_i})
$$

Confidence

- The confidence of proposition X

$$P(\mathbf{x}|\mathbf{e}) \propto \underbrace{\prod_{j=1}^{|\mathcal{C}|} P(\mathbf{e}_{\mathcal{C}_j}|\mathbf{x})}_{P(\mathbf{e}^{\mathcal{C}}|\mathbf{x})} \underbrace{\left[\sum_{all\ \mathcal{P}_{mn}} P(\mathbf{x}|\mathcal{P}_{mn}) \prod_{i=1}^{|\mathcal{P}|} P(\mathcal{P}_i|\mathbf{e}_{\mathcal{P}_i}) \right]}_{P(\mathbf{x}|\mathbf{e}^{\mathcal{P}})}$$

- The probability that node X takes a particular value is equal to the product of two factors
  - The first depends on the child nodes
  - The first depends on the parent nodes



Evidence

- Simple cases
  - e_i directly represent the value of variable
  - Confidence

$$P(x \mid \mathbf{e}) = \frac{P(x, \mathbf{e})}{P(\mathbf{e})} = \alpha\, P(x, \mathbf{e})$$

For a fixed e, α is a constant

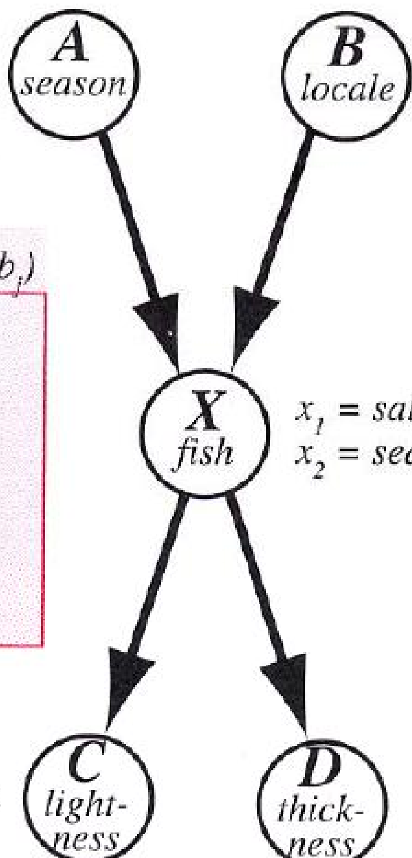
Example One

- fish classification belief net

$P(a)$

$P(a_1)$	$P(a_2)$	$P(a_3)$	$P(a_4)$
0.25	0.25	0.25	0.25

a_1 = winter
a_2 = spring
a_3 = summer
a_4 = autumn



b_1 = north Atlantic
b_2 = south Atlantic

$P(b)$

$P(b_1)$	$P(b_2)$
0.6	0.4

$P(x|a,b)$

	$P(x_1|a_i, b_j)$	$P(x_2|a_i, b_j)$
a_1, b_1	0.5	0.5
a_1, b_2	0.7	0.3
a_2, b_1	0.6	0.4
a_2, b_2	0.8	0.2
a_3, b_1	0.4	0.6
a_3, b_2	0.1	0.9
a_4, b_1	0.2	0.8
a_4, b_2	0.3	0.7

x_1 = salmon
x_2 = sea bass

$P(c|x)$

	$P(c_1|x_k)$	$P(c_2|x_k)$	$P(c_3|x_k)$
x_1	0.6	0.2	0.2
x_2	0.2	0.3	0.5

c_1 = light
c_2 = medium
c_3 = dark

d_1 = wide
d_2 = thin

$P(d|x)$

	$P(d_1|x_k)$	$P(d_2|x_k)$
x_1	0.3	0.7
x_2	0.6	0.4

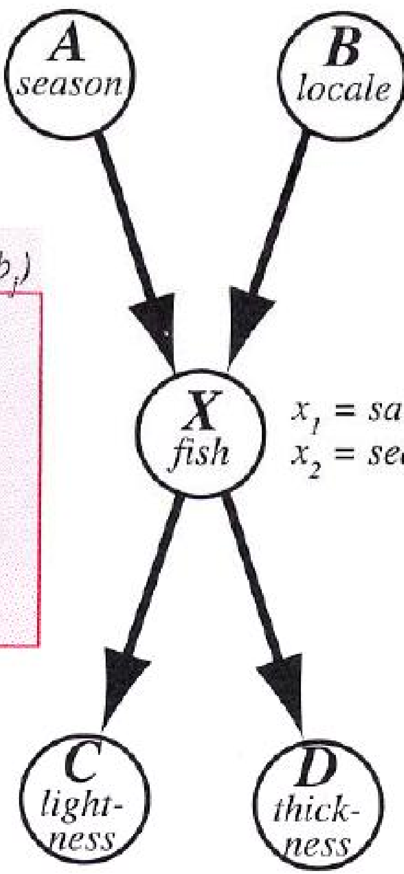
Example One

- Catch a shiny fish in the South Atlantic, salmon or bass?

$P(a)$

$P(a_1)$	$P(a_2)$	$P(a_3)$	$P(a_4)$
0.25	0.25	0.25	0.25

a_1 = winter
a_2 = spring
a_3 = summer
a_4 = autumn



b_1 = north Atlantic
b_2 = south Atlantic

$P(b)$

$P(b_1)$	$P(b_2)$
0.6	0.4

$P(x|a,b)$

	$P(x_1|a_i, b_j)$	$P(x_2|a_i, b_j)$
a_1, b_1	0.5	0.5
a_1, b_2	0.7	0.3
a_2, b_1	0.6	0.4
a_2, b_2	0.8	0.2
a_3, b_1	0.4	0.6
a_3, b_2	0.1	0.9
a_4, b_1	0.2	0.8
a_4, b_2	0.3	0.7

x_1 = salmon
x_2 = sea bass

$P(c|x)$

	$P(c_1|x_k)$	$P(c_2|x_k)$	$P(c_3|x_k)$
x_1	0.6	0.2	0.2
x_2	0.2	0.3	0.5

c_1 = light
c_2 = medium
c_3 = dark

d_1 = wide
d_2 = thin

$P(d|x)$

	$P(d_1|x_k)$	$P(d_2|x_k)$
x_1	0.3	0.7
x_2	0.6	0.4

Example One

- Catch a shiny fish in the South Atlantic, salmon or bass?

e: a is unknown

b_2=South Atlantic

c_1=shinny

d is unknown

First calculate the probability of x_1=salmon

$$
\begin{aligned}
P(x_1 \mid \mathbf{e}) &= \alpha P(x_1, b_2, c_1) \\
&= \alpha \sum_{a,d} (x_1, a, b_2, c_1, d) \\
&= \alpha \sum_{a,d} P(a)P(b_2)P(x_1 \mid a, b_2)P(c_1 \mid x_1)P(d \mid x_1) \\
&= \alpha P(b_2)P(c_1 \mid x_1) \sum_{a} P(a)P(x_1 \mid a, b_2) \sum_{d} P(d \mid x_1) \\
&= \alpha P(b_2)P(c_1 \mid x_1) \times [P(a_1)P(x_1 \mid a_1, b_2) + P(a_2)P(x_1 \mid a_2, b_2) \\
&\quad + P(a_3)P(x_1 \mid a_3, b_2) + P(a_4)P(x_1 \mid a_4, b_2)] \times [P(d_1 \mid x_1) + P(d_2 \mid x_1)] \\
&= \alpha \times 0.4 \times 0.6 \times [0.25 \times 0.7 + 0.25 \times 0.8 + 0.25 \times 0.1 + 0.25 \times 0.3] \times 1.0 \\
&= 0.114\alpha
\end{aligned}
$$

Example One

- Catch a shiny fish in the South Atlantic, salmon or bass?

a is unknown

b_2=South Atlantic

c_1=shinny

d is unknown

$$P(x_2 \mid \mathbf{e}) = 0.042\alpha$$

Then calculate the probability of x_2=bass

- normalization （make $P(x_1 \mid \mathbf{e}) + P(x_2 \mid \mathbf{e}) = 1$ ）

$$P(x_1 \mid \mathbf{e}) = 0.63$$

$$P(x_2 \mid \mathbf{e}) = 0.27$$

Because $P(x_1 \mid \mathbf{e}) > P(x_2 \mid \mathbf{e})$ ， so it's salmon

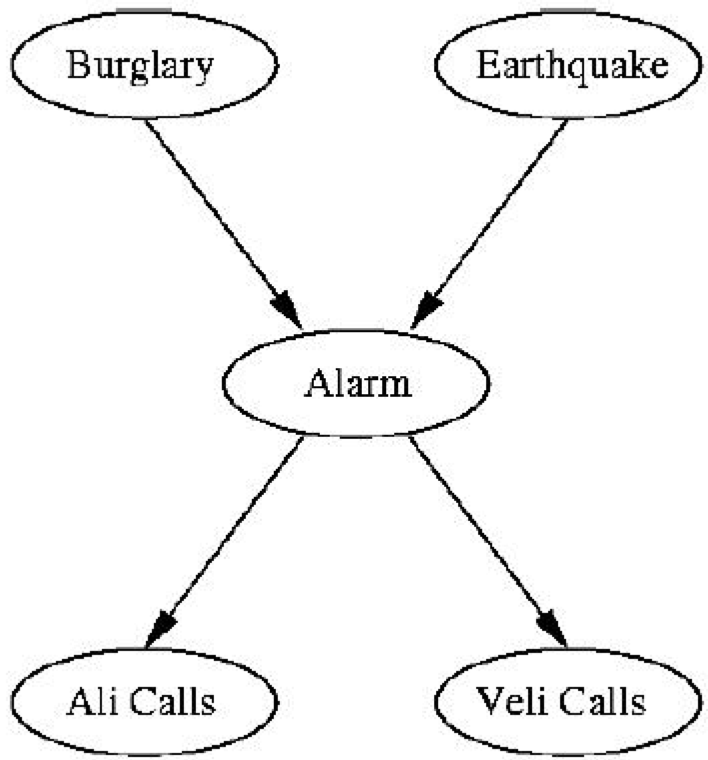
Example Two

- You have installed an anti - theft system in your house
- The system is sensitive to burglary detection, but sometimes earthquakes can trigger alarms
- You have two neighbors: Ali and Veli. When you are not at home, they would call you if they heard the alarm
- Ali would call you when he heard the alarm, but sometimes he would call you because he thought phone ringing was an alarm
- Veli often listens to music at home, so sometimes she doesn't hear the alarm
- Can you estimate the true probability of a real burglary based on which neighbor called you?

Example Two

- Modeling

P(B=T)	P(B=F)
0.001	0.999



P(E=T)	P(E=F)
0.002	0.998

B	E	P(A=T)	P(A=F)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(AC=T)	P(AC=F)
T	0.90	0.10
F	0.05	0.95

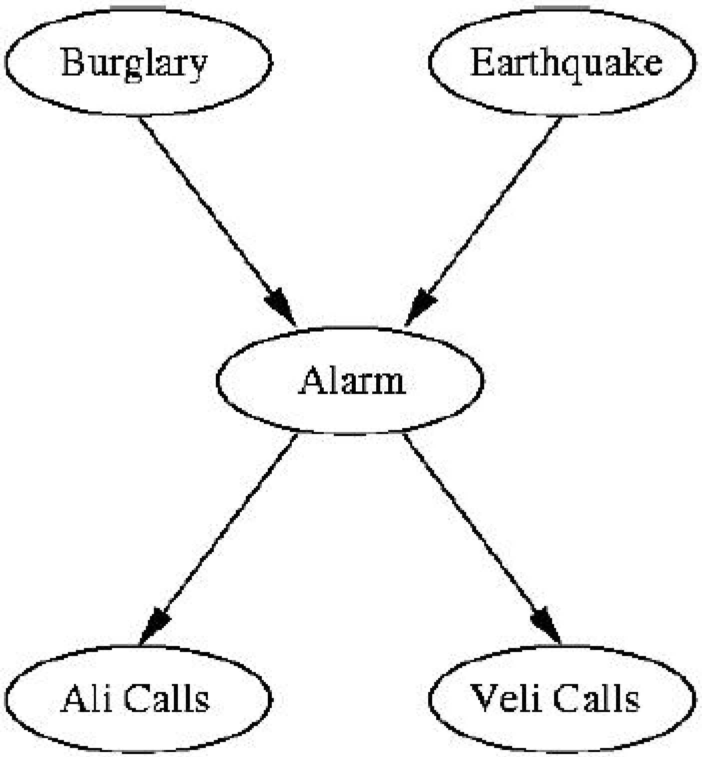
A	P(VC=T)	P(VC=F)
T	0.70	0.30
F	0.01	0.99

Example Two

- The system alarms, but neither burglary nor earthquake occurred, besides both Ali and Veli call you

P(B=T)	P(B=F)
0.001	0.999

P(E=T)	P(E=F)
0.002	0.998



B	E	P(A=T)	P(A=F)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(AC=T)	P(AC=F)
T	0.90	0.10
F	0.05	0.95

A	P(VC=T)	P(VC=F)
T	0.70	0.30
F	0.01	0.99

Example Two

- Calculate the probability of the following event
  - The system alarms, but neither burglary nor earthquake occurred, besides both Ali and Veli call you

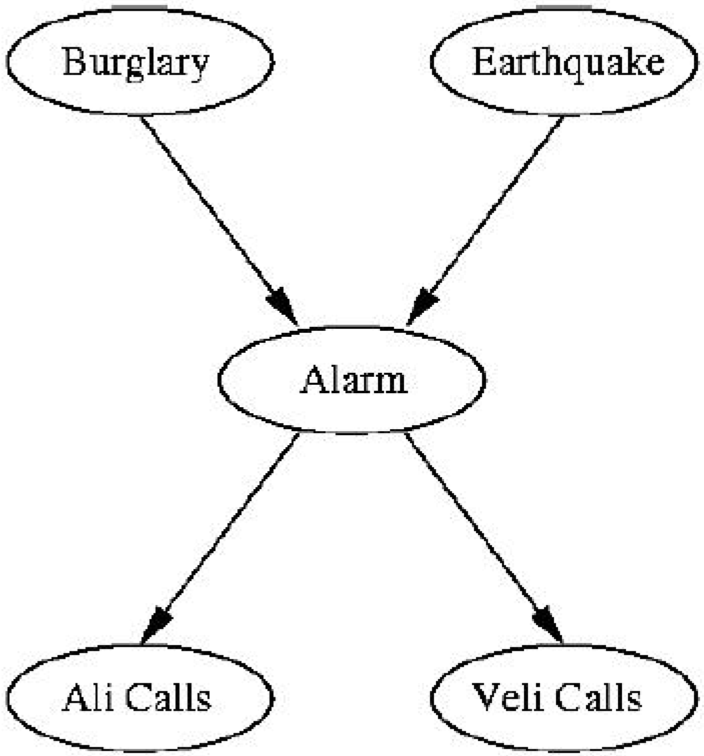
$$
\begin{aligned}
&P(AC, VC, A, \neg B, \neg E) \\
&\quad = P(AC|A)P(VC|A)P(A|\neg B, \neg E)P(\neg B)P(\neg E) \\
&\quad = 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 \\
&\quad = 0.00062
\end{aligned}
$$

Example Two

- If Ali calls you, calculate the confidence that a burglary has occurred

P(B=T)	P(B=F)
0.001	0.999

P(E=T)	P(E=F)
0.002	0.998



B	E	P(A=T)	P(A=F)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(AC=T)	P(AC=F)
T	0.90	0.10
F	0.05	0.95

A	P(VC=T)	P(VC=F)
T	0.70	0.30
F	0.01	0.99

Example Two

- If Ali calls you, calculate the confidence that a burglary has occurred
  - **Method One**

$$P(B \mid AC) = \alpha P(B, AC)$$

$$= \alpha \sum_{vc} \sum_{a} \sum_{e} P(AC \mid a) P(vc \mid a) P(a \mid B, e) P(B) P(e)$$

$$= 0.00084632\alpha$$

$$P(\neg B \mid AC) = \alpha P(\neg B, AC)$$

$$= \alpha \sum_{vc} \sum_{a} \sum_{e} P(AC \mid a) P(vc \mid a) P(a \mid \neg B, e) P(\neg B) P(e)$$

$$= 0.0513\alpha$$

Normalization $P(B \mid AC) = \dfrac{0.00084632\alpha}{0.00084632\alpha + 0.0513\alpha} = 0.0162$

Example Two

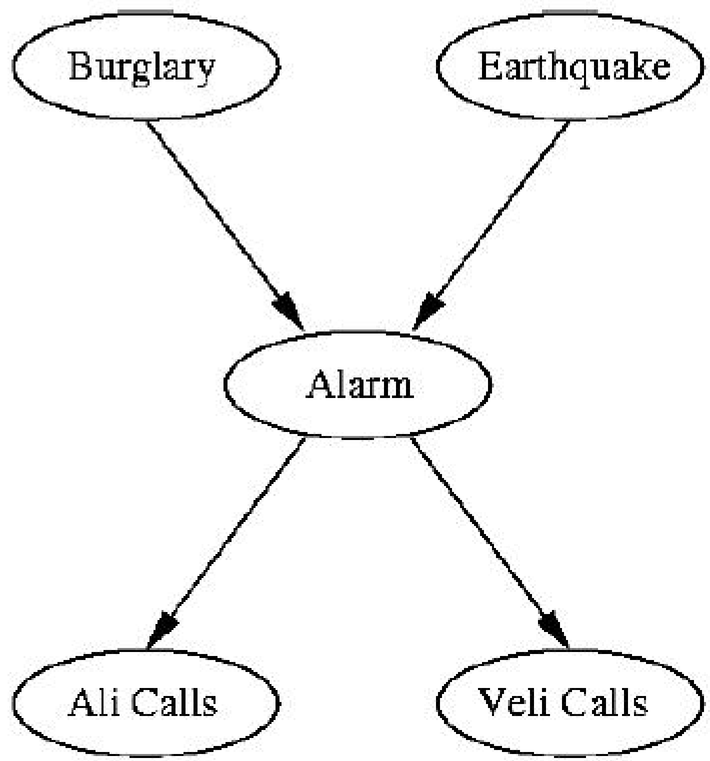
- If Ali calls you, calculate the confidence that a burglary has occurred
  - **Method Two**

$$
\begin{aligned}
P(B|AC) &= \frac{P(B,AC)}{P(AC)} \\
&= \frac{\sum_{vc}\sum_{a}\sum_{e} P(AC|a)P(vc|a)P(a|B,e)P(B)P(e)}{P(B,AC)+P(\neg B,AC)} \\
&= \frac{0.00084632}{0.00084632+0.0513} \\
&= 0.0162
\end{aligned}
$$

Example Two

- If Both Ali and Veli call you, calculate the confidence that a burglary has occurred

P(B=T)	P(B=F)
0.001	0.999



P(E=T)	P(E=F)
0.002	0.998

B	E	P(A=T)	P(A=F)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(AC=T)	P(AC=F)
T	0.90	0.10
F	0.05	0.95

A	P(VC=T)	P(VC=F)
T	0.70	0.30
F	0.01	0.99

Example Two

- If Both Ali and Veli call you, calculate the confidence that a burglary has occurred

$$P(B \mid AC, VC) = \frac{P(B, AC, VC)}{P(AC, VC)}$$

$$= \frac{\sum_a \sum_e P(AC \mid a) P(VC \mid a) P(a \mid B, e) P(B) P(e)}{P(B, AC, VC) + P(\neg B, AC, VC)}$$

$$= 0.29$$

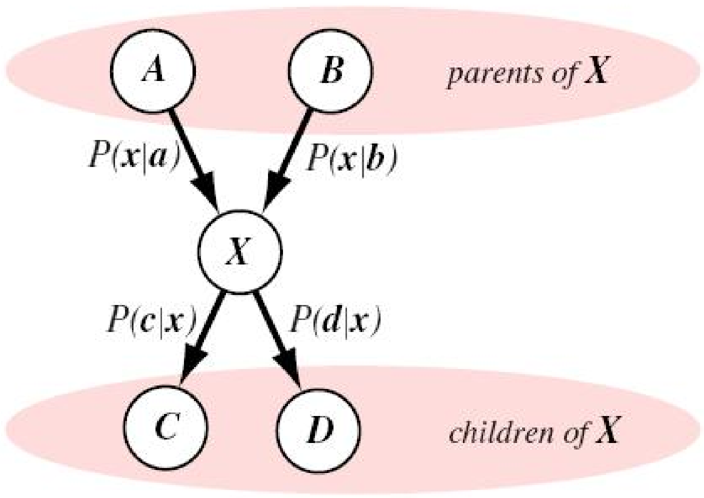
Evidence

- Given the values of variables other than the target variable X, determine the probability of the other variables
- Evidence $\{e_{\mathbf{A}}, e_{\mathbf{B}}, e_{\mathbf{C}}, e_{\mathbf{D}}\}$, where e_i indicate the value of the variable
- For example, fish classification belief net
  - Available evidence $\mathbf{e} = \{e_A, e_B, e_C, \ldots\}$
    - $e_{\mathbf{A}}$: winter now
      $P(a_1|e_{\mathbf{A}}) = 1 \quad P(a_i|e_{\mathbf{A}}) = 0 \text{ for } i = 2, 3, 4$
    - $e_{\mathbf{B}}$: fishermen prefer the South Atlantic
      $P(b_1|e_{\mathbf{B}}) = 0.2 \qquad P(b_2|e_{\mathbf{B}}) = 0.8$
    - $e_{\mathbf{C}}$: the fish has a lighter sheen
      $P(e_{\mathbf{C}}|c_1) = 1 \qquad P(e_{\mathbf{C}}|c_2) = 0.5 \qquad P(e_{\mathbf{C}}|c_3) = 0$
    - $e_{\mathbf{D}}$: the width cannot be measured because of occlusion
      $P(e_{\mathbf{D}}|d_1) = P(e_{\mathbf{D}}|d_2)$

Pay attention to the position of e_i!

Confidence

- Consider a certain node X
- The set of nodes before X is called the parent node P of X and the set of nodes after X is called the child node C of X
- For example：
  - parent node of X：{A，B}
  - child node of X ：{C，D}



- When estimating the probability of X, the parent node and the child node of X should be treated differently
  - Evidence **e**：Values of variables at nodes other than X
  - Given e, the Confidence Belief of x = (x1, x2...)

$$P(\mathbf{x}|\mathbf{e}) \propto P(\mathbf{e}^{\mathcal{C}}|\mathbf{x})P(\mathbf{x}|\mathbf{e}^{\mathcal{P}})$$

  - Must be normalized so that the sum of the probabilities of all values of x is 1

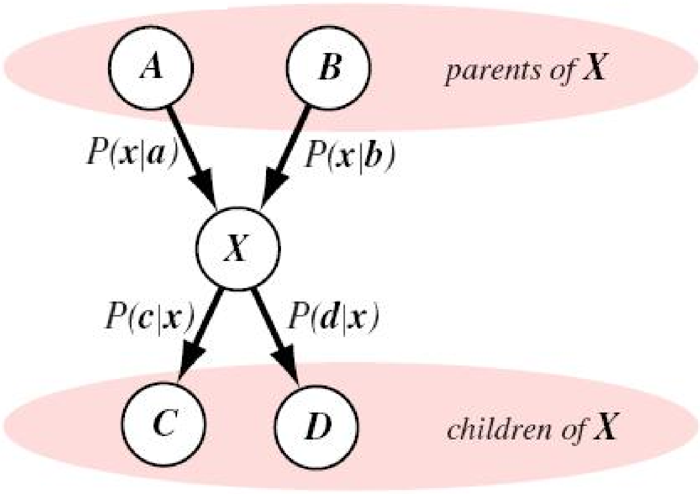
Confidence

- For child node of X
  - Suppose there are no connections between child nodes

$$
\begin{aligned}
P(\mathbf{e}^{\mathcal{C}}|\mathbf{x}) &= P(\mathbf{e}_{\mathcal{C}_1}, \mathbf{e}_{\mathcal{C}_2}, ..., \mathbf{e}_{\mathcal{C}_{|\mathcal{C}|}}|\mathbf{x}) \\
&= P(\mathbf{e}_{\mathcal{C}_1}|\mathbf{x})P(\mathbf{e}_{\mathcal{C}_2}|\mathbf{x}) \cdots P(\mathbf{e}_{\mathcal{C}_{|\mathcal{C}|}}|\mathbf{x}) \\
&= \prod_{j=1}^{|\mathcal{C}|} P(\mathbf{e}_{\mathcal{C}_j}|\mathbf{x}),
\end{aligned}
$$

  - For exmple：

$$P(\mathbf{e}_{\mathbf{C}}, \mathbf{e}_{\mathbf{D}}|\mathbf{x}) = P(\mathbf{e}_{\mathbf{C}}|\mathbf{x})P(\mathbf{e}_{\mathbf{D}}|\mathbf{x})$$

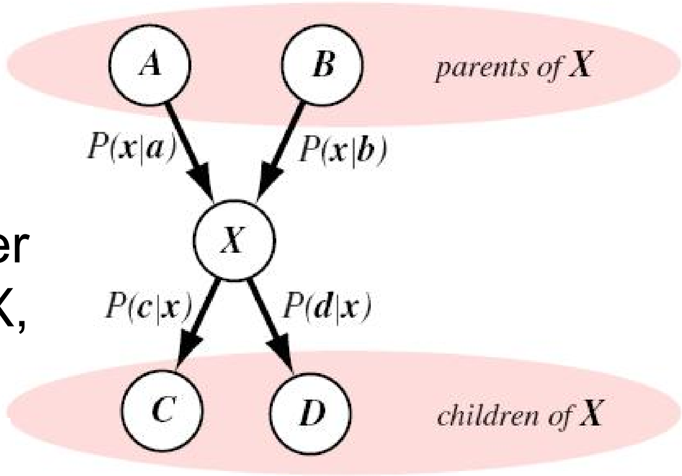


Confidence

- For parent node of X
  - Suppose there are no connections between parent nodes

$$
\begin{aligned}
P(\mathrm{x}|\mathrm{e}^{\mathcal{P}}) &= P(\mathrm{x}|\mathrm{e}_{\mathcal{P}_1}, \mathrm{e}_{\mathcal{P}_2}, ..., \mathrm{e}_{\mathcal{P}_{|\mathcal{P}|}}) \\
&= \sum_{all\ i,j,...,k} P(\mathrm{x}|\mathcal{P}_{1i}, \mathcal{P}_{2j}, ..., \mathcal{P}_{|\mathcal{P}|k}) P(\mathcal{P}_{1i}, \mathcal{P}_{2j}, ..., \mathcal{P}_{|\mathcal{P}|k}|\mathrm{e}_{\mathcal{P}_1}, ..., \mathrm{e}_{\mathcal{P}_{|\mathcal{P}|}}) \\
&= \sum_{all\ i,j,...,k} P(\mathrm{x}|\mathcal{P}_{1i}, \mathcal{P}_{2j}, ..., \mathcal{P}_{|\mathcal{P}|k}) P(\mathcal{P}_{1i}|\mathrm{e}_{\mathcal{P}_1}) \cdots P(\mathcal{P}_{|\mathcal{P}|k}|\mathrm{e}_{\mathcal{P}_{|\mathcal{P}|k}}),
\end{aligned}
$$

  - $\mathcal{P}_{mn}$ represents the value of the parent node $\mathcal{P}_m$ in state n
  - Ignore node interdependencies other than the parent and child nodes of X, and simplify above equation



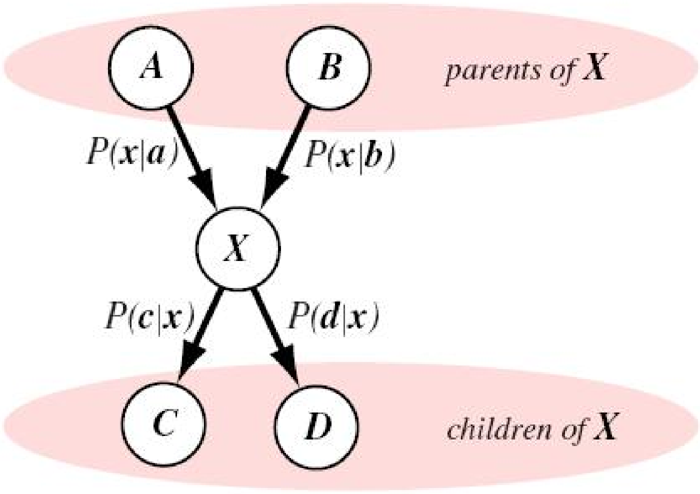
$$
P(\mathrm{x}|\mathrm{e}^{\mathcal{P}}) = \sum_{all\ \mathcal{P}_{mn}} P(\mathrm{x}|\mathcal{P}_{mn}) \prod_{i=1}^{|\mathcal{P}|} P(\mathcal{P}_i|\mathrm{e}_{\mathcal{P}_i})
$$

Confidence

- The confidence of proposition X

$$P(\mathbf{x}|\mathbf{e}) \propto \underbrace{\prod_{j=1}^{|\mathcal{C}|} P(\mathbf{e}_{\mathcal{C}_j}|\mathbf{x})}_{P(\mathbf{e}^{\mathcal{C}}|\mathbf{x})} \underbrace{\left[\sum_{all\ \mathcal{P}_{mn}} P(\mathbf{x}|\mathcal{P}_{mn}) \prod_{i=1}^{|\mathcal{P}|} P(\mathcal{P}_i|\mathbf{e}_{\mathcal{P}_i}) \right]}_{P(\mathbf{x}|\mathbf{e}^{\mathcal{P}})}$$

- The probability that node X takes a particular value is equal to the product of two factors
  - The first depends on the child nodes
  - The first depends on the parent nodes

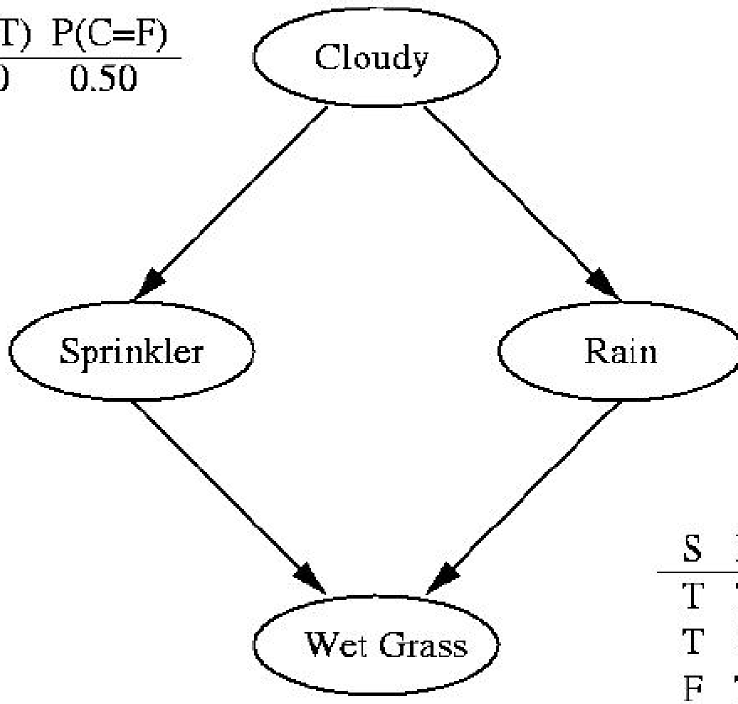


Example Three

- The grass can get wet for two reasons: the sprinklers have been turned on, or it has rained
- If it's cloudy, it's more likely to rain than it is sunny
- If it is cloudy, it’s less likely to turn on the sprinklers
- Suppose it's equally likely to be cloudy or sunny

Example Three

- Modeling



P(C=T)	P(C=F)
0.50	0.50

C	P(S=T)	P(S=F)
T	0.10	0.90
F	0.50	0.50

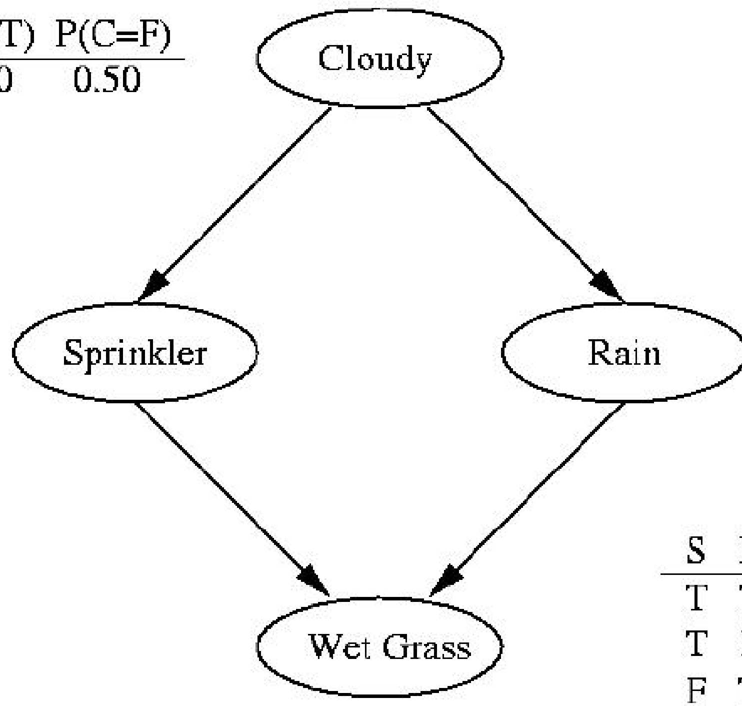
C	P(R=T)	P(R=F)
T	0.80	0.20
F	0.20	0.80

S	R	P(W=T)	P(W=F)
T	T	0.99	0.01
T	F	0.90	0.10
F	T	0.90	0.10
F	F	0.00	1.00

Example Three

- If you see that the grass is wet, which reason is more likely to be the sprinklers or the rain?

P(C=T)	P(C=F)
0.50	0.50



C	P(S=T)	P(S=F)
T	0.10	0.90
F	0.50	0.50

C	P(R=T)	P(R=F)
T	0.80	0.20
F	0.20	0.80

S	R	P(W=T)	P(W=F)
T	T	0.99	0.01
T	F	0.90	0.10
F	T	0.90	0.10
F	F	0.00	1.00

Example Three

- If you see that the grass is wet, which reason is more likely to be the sprinklers or the rain?

$$P(S|W) = \frac{P(S,W)}{P(W)} = \frac{\sum_c \sum_r P(W|S,r)P(S|c)P(r|c)P(c)}{P(S,W) + P(\neg S,W)}$$

$$= \frac{0.2781}{0.2781 + 0.369} = 0.430$$

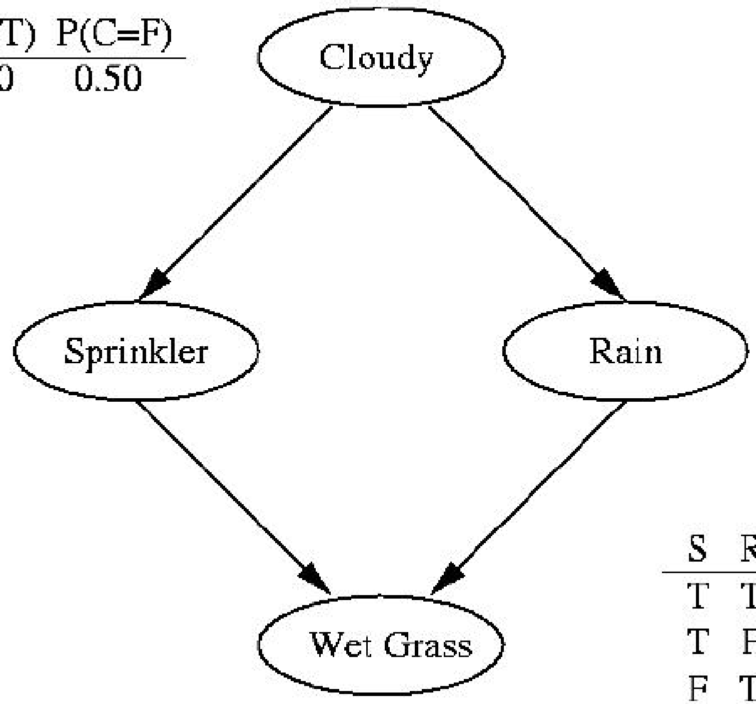
$$P(R|W) = \frac{P(R,W)}{P(W)} = \frac{\sum_c \sum_s P(W|s,R)P(s|c)P(R|c)P(c)}{P(R,W) + P(\neg R,W)}$$

$$= \frac{0.4581}{0.6471} = 0.708$$

Because $P(S|W) < P(R|W)$**, So it is more likely that the grass would be wet by rain**

Example Three

- What if the grass is wet and the weather is sunny?



P(C=T)	P(C=F)
0.50	0.50

C	P(S=T)	P(S=F)
T	0.10	0.90
F	0.50	0.50

C	P(R=T)	P(R=F)
T	0.80	0.20
F	0.20	0.80

S	R	P(W=T)	P(W=F)
T	T	0.99	0.01
T	F	0.90	0.10
F	T	0.90	0.10
F	F	0.00	1.00

Example Three

- What if the grass is wet and the weather is sunny?

$$P(S \mid W, \neg C) = \frac{P(S, W, \neg C)}{P(W, \neg C)} = \frac{P(S \mid \neg C) P(\neg C) \sum_r P(W \mid S, r) P(r \mid \neg C)}{P(S, W, \neg C) + P(\neg S, W, \neg C)}$$

$$= \frac{0.2295}{0.2295 + 0.045} = 0.836$$

$$P(R \mid W, \neg C) = \frac{P(R, W, \neg C)}{P(W, \neg C)} = \frac{P(R \mid \neg C) P(\neg C) \sum_s P(W \mid s, R) P(s \mid \neg C)}{P(R, W, \neg C) + P(\neg R, W, \neg C)}$$

$$= \frac{0.0945}{0.2745} = 0.344$$

Because $P(S \mid W, \neg C) > P(R \mid W, \neg C)$**,**
So it is more likely that the grass would be wet by sprinkling

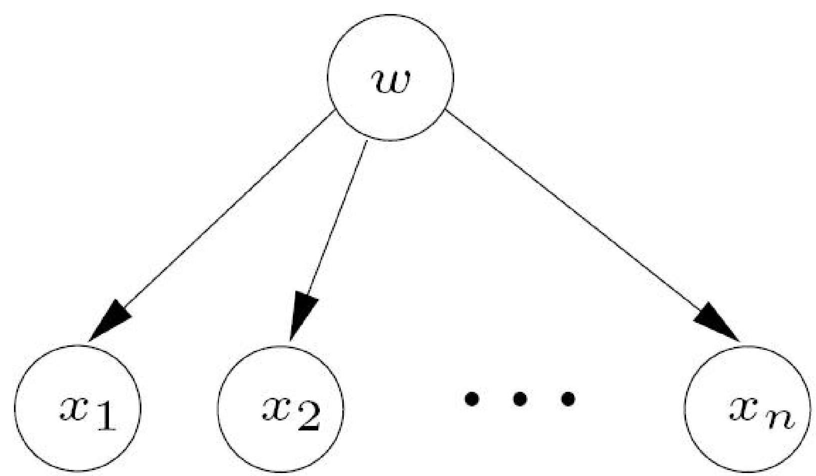
Naïve Bayes Rule

- When the dependencies between features are unknown, it is often assumed that each feature is conditionally independent under a given category condition

$$p(\mathbf{x} \mid \omega_k) = \prod_{i=1}^{d} p(x_i \mid \omega_k)$$

- assumption is called naïve Bayes rule (朴素贝叶斯规则) or idiot Bayes rule (傻瓜贝叶斯规则)

- Naïve Bayes belief net



Part 3 Expected Maximization (EM) algorithm

Missing Feature

- Suppose there is a Bayesian classifier based on the eigenvector $\mathbf{x}$, in which a part of the feature $\mathbf{x}_g$ is visible in each $\mathbf{x}$ and the rest of the feature $\mathbf{x}_b$ is missing. The missing features may be different in different samples
- How to make decisions?
  - Method One
    - simply throw away the sample containing the missing value
  - Method Two
    - replace $\mathbf{x}_b$ with some other sample mean $\overline{\mathbf{x}}_b$ known to have this feature, that is $\mathbf{x} = (\mathbf{x}_g \overline{\mathbf{x}}_b)$

Expectation Maximization

- **Mehod Three**
  - By extending the maximum likelihood method, the model parameters can be learned in the case of missing values in the training set
  - **expectation-maximization, EM** （**期望最大化**）Algorithm can estimate likelihood function based on existing data recursively
- Two main applications of EM algorithm
  - Learn when the data is incomplete or has missing values
  - When the likelihood equation is difficult to solve directly, initialize some unknown parameters to simplify the problem

Expectation Maximization

- Suppose sample **x** obeys a certain distribution $p(\mathbf{x} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the unknown parameter vector
- Sample set $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, where $\mathbf{x}_k = \{\mathbf{x}_{kg}, \mathbf{x}_{kb}\}$, $\mathbf{x}_{kg}$ is the complete (or good) part and $\mathbf{x}_{kb}$ is the missing (or damaged) part
- represent the set of $\mathbf{x}_{kg}$ and $\mathbf{x}_{kb}$ separately $\mathcal{D} = \mathcal{D}_g \cup \mathcal{D}_b$
- Given some assumption θ^i (not necessarily accurate) for θ , calculate the expectation of the log-likelihood function under a distribution determined by θ^i about missing features, and get a function of θ

$$Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i) = \mathcal{E}_{\mathcal{D}_b}[\ln p(\mathcal{D}_g, \mathcal{D}_b;\ \boldsymbol{\theta}) | \mathcal{D}_g;\ \boldsymbol{\theta}^i]$$

$$= \int_{-\infty}^{+\infty} \ln p(D_g, D_b \mid \theta) p(D_b \mid \theta^i; D_g) dD_b$$

Expectation Maximization

$$Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i) = \mathcal{E}_{\mathcal{D}_b}[\ln p(\mathcal{D}_g, \mathcal{D}_b;\ \boldsymbol{\theta})|\mathcal{D}_g;\ \boldsymbol{\theta}^i]$$
$$= \int_{-\infty}^{+\infty} \ln p(D_g, D_b \mid \theta) p(D_b \mid \theta^i; D_g) dD_b$$

- Meaning: the parameter vector $\boldsymbol{\theta}^i$ is the current estimate of the parameter and $\boldsymbol{\theta}$ is a candidate parameter vector to improve the current estimate. For each $\boldsymbol{\theta}$, the log-likelihood function of the training set can be computed. Due to the presence of missing values $\mathcal{D}_b$, the likelihood function needs to marginalize $\mathcal{D}_b$, and the distribution of $\mathcal{D}_b$ in the edge integral is determined by the current estimate $\boldsymbol{\theta}^i$

Expectation Maximization

- **expectation-maximization(EM) algorithm**

begin **initialize** $\boldsymbol{\theta}^0, T, i = 0$

  do $i \leftarrow i + 1$

    E step : compute $Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i)$

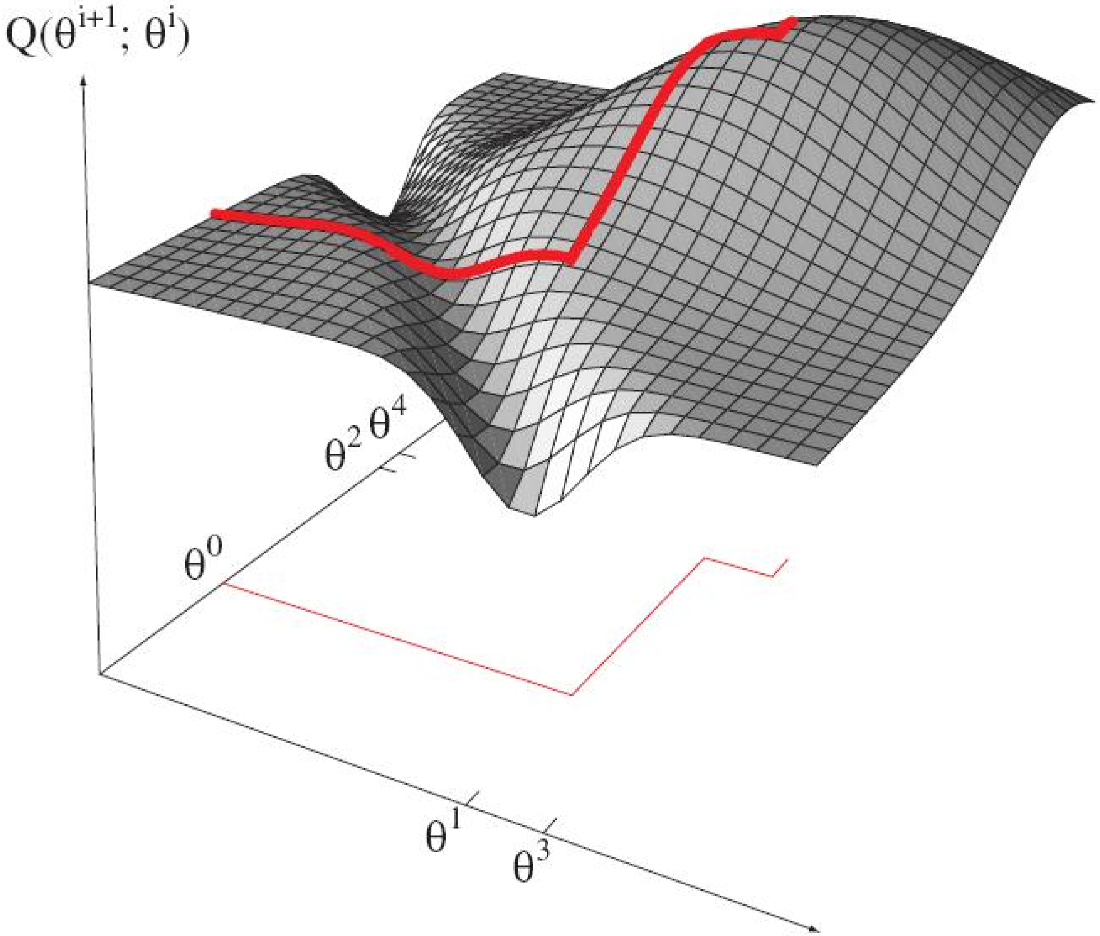
    M step : $\boldsymbol{\theta}^{i+1} \leftarrow \arg\max_{\theta} Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i)$

  until $Q(\boldsymbol{\theta}^{i+1};\ \boldsymbol{\theta}^i) - Q(\boldsymbol{\theta}^i;\ \boldsymbol{\theta}^{i-1}) \leq T$

 return $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{i+1}$

end

Expectation Maximization



Example

- EM algorithm for two-dimensional normal distribution
  - Dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} = \left\{ \binom{0}{2}, \binom{1}{0}, \binom{2}{2}, \binom{*}{4} \right\}$
  - Probabilistic Model
    - two-dimensional normal distribution, $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$
  - Target: To estimate the parameter vectors of a normal distribution

$$\boldsymbol{\theta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}$$

Example

- Solution
  - initialization
  - The mean is at the origin, the covariance matrix is the identity matrix, that is

$$\theta^0 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

  - Calculate θ^1
    - **Step E**

Example

$$
\begin{aligned}
Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^0) &= \mathcal{E}_{x_{41}}[\ln p(\mathbf{x}_g, \mathbf{x}_b;\ \boldsymbol{\theta}|\boldsymbol{\theta}^0;\ \mathcal{D}_g)] \\
&= \int_{-\infty}^{\infty} \left[\sum_{k=1}^{3} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) + \ln p(\mathbf{x}_4|\boldsymbol{\theta}) \right] p(x_{41}|\boldsymbol{\theta}^0;\ x_{42}=4)\ dx_{41} \\
&= \sum_{k=1}^{3} [\ln p(\mathbf{x}_k|\boldsymbol{\theta})] + \int_{-\infty}^{\infty} \ln p\left(\left(\begin{matrix} x_{41} \\ 4 \end{matrix} \right) \middle| \boldsymbol{\theta} \right) \frac{p\left(\left(\begin{smallmatrix} x_{41} \\ 4 \end{smallmatrix} \right) \middle| \boldsymbol{\theta}^0 \right)}{\underbrace{\left(\int_{-\infty}^{\infty} p\left(\left(\begin{matrix} x'_{41} \\ 4 \end{matrix} \right) \middle| \boldsymbol{\theta}^0 \right) dx'_{41} \right)}_{\equiv K}} dx_{41} \\
&= \sum_{k=1}^{3} [\ln p(\mathbf{x}_k|\boldsymbol{\theta})] + \frac{1}{K} \int_{-\infty}^{\infty} \ln p\left(\left(\begin{matrix} x_{41} \\ 4 \end{matrix} \right) \middle| \boldsymbol{\theta} \right) \frac{1}{2\pi \left| \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix} \right) \right|} \exp\left[-\frac{1}{2}(x_{41}^2 + 4^2) \right] dx_{41} \\
&= \sum_{k=1}^{3} [\ln p(\mathbf{x}_k|\boldsymbol{\theta})] - \frac{1+\mu_1^2}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2} - \ln\ (2\pi\sigma_1\sigma_2).
\end{aligned}
$$

Example

- **Step M**

$$\nabla_{\theta} Q(\theta \mid \theta^{0}) = 0$$

$$\boldsymbol{\theta}^1 = \begin{pmatrix} 0.75 \\ 2.0 \\ 0.938 \\ 2.0 \end{pmatrix}$$

- Iterate Step E and Step M, and after 3 iterations θ converges to

$$\boldsymbol{\mu} = \begin{bmatrix} 1.0 \\ 2.0 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 0.667 & 0 \\ 0 & 2.0 \end{bmatrix}$$

Example