Probability Space:

(Ω, E, P)

- Ω: sample space
- E: Event
- P: distribution $P: E \to \Omega$

Random Variables: X (upper case)

- $X: \Omega \to S$.
- set of values X can take: $Val(X)$
- $P(x) \Leftrightarrow P(X = x)$
- x^i: a special value of X

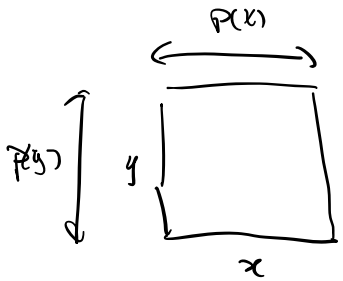
Prob Review:

1. Joint Prob: 2 or more events: $X, Y, Z, \ldots$

$$P(X = x, Y = y)$$

2. Maginal:

$$p(x) = \int p(x,y)\,dy \quad , \quad p(y) = \int p(x,y)\,dx$$



3. Conditional:

- $p(x|Y=y)$:= prob of X given $Y=y$.

- Calculation:

$$p(x|Y=y) = \frac{p(x, Y=y)}{p(Y=y)}$$

$$\Rightarrow p(x|y) = \frac{p(x,y)}{p(y)}$$

$$\Rightarrow p(w,x,y,z) = p(w,x,y|z) \cdot p(z)$$
$$= p(w,x|y,z) \cdot p(y|z) \cdot p(z)$$
$$= p(w|x,y,z) \cdot p(x|y,z)\, p(y|z)\, p(z)$$ ——— chain rule

- Bayes: $p(y|x) = \dfrac{p(x|y)\,p(y)}{\int p(x|y)\,p(y)\,dy}$

likelihood ← $p(x|y)$; $p(y)$ → prior

$p(y|x)$ ↓ posterior

$\int p(x|y)\,p(y)\,dy$ ↳ Evidence (constant, normalizer)

(y is latent variable, x is observation)

4. independence: $p(x|y) = p(x)$, $p(y|x) = p(y)$

$\Rightarrow p(x,y) = p(x)p(y)$

5. Expectation: $E[f(x)] = \sum_x f(x)p(x)$

$E[f(x)] = \int_x f(x)p(x)dx$

Rules: ① $E(c) = c$

② $E(c \cdot f(x)) = c \cdot E(f(x))$

③ $E(f(x) + g(x)) = E(f(x)) + E(g(x))$

④ $E(f(x) \cdot g(y)) = E(f(x)) \cdot E(g(y)) \Leftrightarrow X, Y$ independent

$\Downarrow$

$\int f(x)dx \cdot \int g(y)dy$

6. common distribution:

① Bernoulli: $p(x|\lambda) = \begin{cases} 1-\lambda & , x=0 \\ \lambda & , x=1 \end{cases}$

denote as: $p(x) = Bern_x[\lambda]$

② Gaussian:

I. single R.V., 2 para:

$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

denote as: $Norm_x[\mu, \sigma^2]$

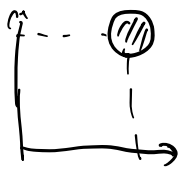
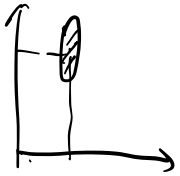
II. D-dimensional R.V. X, para: $\mu \in R^D$, $\Sigma \in R_+^{D \times D}$ → symmetrical, positive covariance matrix

$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$

denote as: $Norm_x[\mu, \Sigma]$

Type of covariance: spherical, diagonal, full

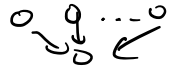
$\Sigma_{sphere} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$ $\Sigma_{diag} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ $\Sigma_{full} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$


Summary: (FOU AI - probabilitic Reasoning and decision making).

① Probability: { Marginalize., independence (marginal. conditional)
product
Bayesian

② Bayes Net: (represent relationship between variables)

- inference:
  - calculate prob with full joint
  - variable elimination. ← { multiply, sum, conditioning } — polytree.
    - ↳ order: { leaves, non-query, non-evidence.
  - independence: { d-seperation. { 1. $X \to E \to Y$, 2. $X \leftarrow E \to Y$, 3. $X \to O \leftarrow Y$ (all Os are non-evidence) } markov blanket
  - Noisy-OR: (n para vs 2^n para)  $P(Y|X_1 \cdots X_n) = 1 - \prod_{i=1}^{n} (1-p_i)^{x_i}$
- learning:
  - MLE for complete data
  - EM for incomplete data (EM for Noisy-OR)
- special BN:
  - naive Bayes
  - HMM (forward Algorithm)

Fudan Summer:

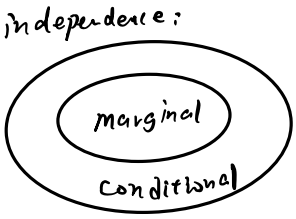
L3:

· Conditional Independence:

A, B conditional independent $\Leftrightarrow$ $p(A|BC) = p(A|C)$. $P(B|AC) = p(B|C)$

then we can get: $P(AB|C) = P(A|C) \cdot P(B|C)$

$P(ABC) = P(C) \cdot P(A|C) \cdot P(B|C)$



· marginally Independent:

A, B marginally independent $\Leftrightarrow$ $P(A|B) = P(A)$. $P(B|A) = P(B)$

· CPT: conditional prob table

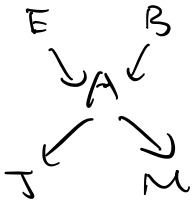
☆ Bayes Net:

BN(def): A DAG with: Node —— Variables; edge —— conditional dependence; CPT at each node —— how a node depends on its parents

E: Earthquake happens
B: Burglar breaks in
A: alarm goes off
J: Jamal calls
M: Maya calls

↳ Binary Events

BN: (dependency)



Can CPT:
$p(E)$, $p(B)$
$p(A|EB)$
$p(J|A)$, $p(M|A)$

1. Compare: $p(B=1)$, $p(B=1|A=1)$, $p(B=1|A=1, E=1)$

$p(B=1|A=1) > p(B=1|A=1, E=1) > p(B=1)$

Intuitive Reason: $P(B=1) < P(B=1|A=1, E=1) < P(B=1|A=1)$

$p(B=1) < p(B=1|A=1)$

I.e. "I generally believe there is a very low chance that a burglar breaks in. But if I get notified that my alarm is going off, then I think the alarm is probably caused by the burglar and my belief that a burglar broke in goes up. However, if I then hear about an earthquake, I no longer am as fearful that a burglar broke in, but I'm still more worried than I would have been if the alarm hadn't gone off at all"

$p(B=1|A=1) > p(B=1|A=1, E=1)$ $p(B=1|A=1, E=1) > p(B=1)$

2. Calculate: $p(B=1|A=1)$. $p(B=1|A=1, E=1)$

☆: E, B: marginal independent . ie. $p(E|B) = p(E)$. $p(B|E) = p(B)$

① $p(B=1|A=1) = \frac{P(A=1|B=1)\cdot P(B=1)}{P(A=1)} = \frac{\left(P(A=1,E=1|B=1)+P(A=1,E=0|B=1)\right)\cdot P(B=1)}{\sum_{i,j} P(A=1,E=e_i,B=b_j)}$

$= \frac{\left(P(A=1|E=1,B=1)\cdot P(E=1|B=1)+P(A=1|E=0,B=1)\cdot P(E=0|B=1)\right)\cdot P(B=1)}{\sum_{i,j} P(A=1,E=e_i,B=b_j)}$

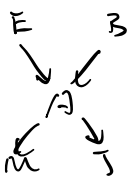
marginal independent

$= \frac{\left(P(A=1|E=1,B=1)\cdot P(E=1)+P(A=1|E=0,B=1)\cdot P(E=0)\right)\cdot P(B=1)}{\sum_{i,j} P(A=1,E=e_i,B=b_j)}$

② $p(B=1|A=1,E=1) = \frac{P(B=1,A=1,E=1)}{P(A=1,E=1)} = \frac{P(A=1|E=1,B=1)\cdot P(E=1)\cdot P(B=1)}{\sum_i P(A=1,B=b_i,E=1)}$

3. compare: $P(A=1)$, $P(A=1|J=1)$, $P(A=1|J=1,M=1)$

$P(A=1|J=1,M=1) > P(A=1|J=1) > P(A=1)$



4. calculate: $P(A=1|J=1)$, $P(A=1|J=1,M=1)$

条件: J, M are conditional independent from each other and from E, B given A

ie: $P(J|A,B,E,M) = P(J|A)$

① $p(A=1|J=1) = \frac{P(J=1|A=1)\cdot P(A=1)}{P(J=1)} = \frac{\cdot\cdot}{P(J=1|A=1)\cdot P(A=1)+P(J=1|A=0)\cdot P(A=0)}$

② $P(A=1|J=1,M=1) = \frac{P(A=1,J=1,M=1)}{P(J=1,M=1)} = \frac{\sum_{i,j} P(A=1,J=1,M=1,B=b_i,E=e_j)}{P(M=1|J=1)\cdot P(J=1)}$

$= \frac{}{\left(P(M=1,A=1|J=1)+P(M=1,A=0|J=1)\right)\cdot}$

$= \frac{}{\left(P(M=1|A=1)\cdot P(A=1|J=1)+P(M=1|A=0)\cdot P(A=0|J=1)\right)\cdot}$

5. calculate: $P(A=1|J=1,B=1) = \frac{P(A=1,J=1,B=1)}{P(J=1,B=1,A=1)+P(J=1,B=1,A=0)}$

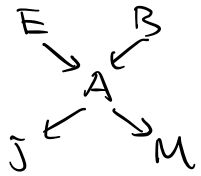
$= \frac{P(J=1|B=1,A=1)\cdot P(A=1|B=1)\cdot P(B=1)}{P(J=1|B=1,A=1)\cdot P(A=1|B=1)\cdot P(B=1)+P(J=1|B=1,A=0)\,P(A=0|B=1)P(B=1)}$

$= \frac{P(J=1|A=1)\cdot P(A=1|B=1)\cdot P(B=1)}{P(J=1|A=1)\cdot P(A=1|B=1)\cdot P(B=1)+P(J=1|A=0)\cdot P(A=0|B=1)\cdot P(B=1)}$

6. Full Joint distribution: $P(A, M, J, B, E)$

↳ which is the most easiest prob to compute with BN.

Eg:



$$P(A, M, J, B, E) = P(M|A) \cdot P(J|A) \cdot P(A|BE) \cdot P(B) \cdot P(E)$$

Complexity: (disadvantage)

for $\{X_1 \dots X_n\}$ n variables.

① # of values need to specify Full Joint: 2^n

② # of values needed to specify the BN: $O(n \cdot 2^k)$

(k: max # of parents any node has)

7. BN. BN inference.

$$P(X_1 \dots X_n) = P(x_n | X_{n-1} \cdot X_{n-2} \cdots X_1) \cdot P(X_{n-1} | X_{n-2} \cdots X_1) \cdots P(X_2 | X_1) \cdot P(X_1)$$

$$= \prod_{i=1}^{n} P(X_i | X_{i-1} \cdots X_1)$$

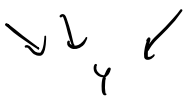
$$= \prod_{i=1}^{n} P(x_i | \text{parent}(x_i)), \quad \text{parent}(x_i) \in \{X_1, \dots X_{i-1}\}$$

8. How to build a BN?

1. choose R.V. $\{X_1 \dots X_n\}$
2. choose an ordering
3. while there are varinbles left:
   - I. add node X_i to BN
   - II. Set parents(X_i): parents ;
     non-parents: conditional independent to X_i
   - III. Define $P(X_i | \text{parent}(X_i))$ CPT.

9. Better Ordering: Root events first (reason: reduce complexity of BN.)

10. Model: X_1 X_2 ... X_k



I. Representing CPTs:

advantage: easy to read

disadvantage: size gives

X_1 X_2 ... X_k	$P(Y=1)$
0 0 .. 0	$P(Y=1 \mid X_1=0, \dots X_k=0)$
0 0 ... 1	$P(Y=1 \mid X_1=1, \dots X_k=0)$
⋮ ⋮ ⋮	⋮

exponentially	1	1 ..	1	$P(Y=1 \mid X_1=1, \dots X_k=1)$

D. Representing OR logic: make $P(Y=1)$ deterministic of its parents.

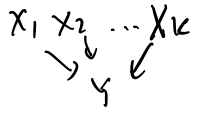
$$P(Y=1 \mid X_1 \cdots X_k) = OR(X_1 \cdots X_k)$$

$$\hookrightarrow OR(X_1 \cdots X_k) = \begin{cases} 1, & \text{if any } X_i=1 \\ 0, & \text{otherwise.} \end{cases}$$

advantage: save space (just calculation)

disadvantage: deterministic

Noisy-OR:



$$P(Y=1 \mid X_1 \cdots X_k) = 1 - \prod_{i=1}^{k} (1-p_i)^{X_i}$$ ——— a k-para model.

$$\hookrightarrow p_i \overset{\text{def}}{=} p(X_i=1, \text{others}=0)$$

Intuition:

- all parents off $\Rightarrow$ child is off certainly
- more parents on $\Rightarrow$ more likely child is on
- all parents on doesn't ensure that child is on (but very likely it's)

11. 3 equation for X, Y conditional indepence given E:

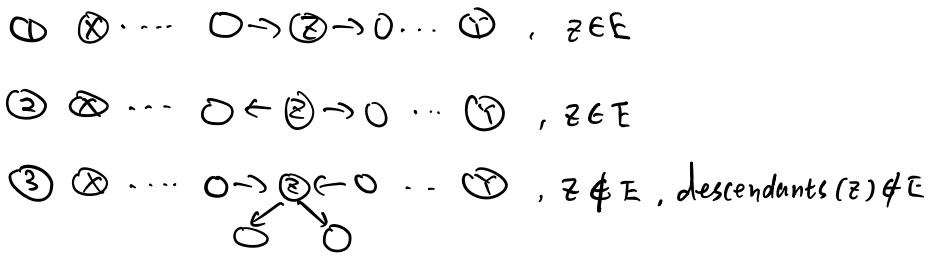
$$\begin{cases} P(X \mid Y, E) = P(X \mid E) \\ P(Y \mid X, E) = P(Y \mid E) \\ P(X, Y \mid E) = P(X \mid E) \cdot P(Y \mid E) \end{cases}$$

12. BN and conditional indepence:

X, Y, E — sets and disjoint. when X, Y independent given E?

Thm: $P(X, Y \mid E) = P(X \mid E) \cdot P(Y \mid E)$ $\Longleftrightarrow$ every undirected path from a node in X to a node in Y is d-seperated (blocked) by E

D-Seperated (def): A path π is d-seperated if $\exists z \in \pi$ holds 1 of these 3 conditions:



Checking Steps:

I. list X, Y, E

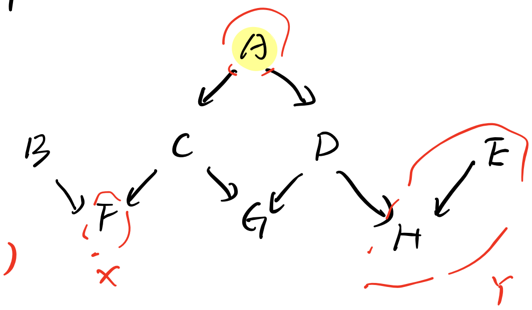
II. list all path $X - Y$, regardless directions

III. put back the directions, check if all paths are d-sep

E

ex. $P(F,H \mid A,E) = P(F \mid A)\,P(H \mid A,E)$?

$X=\{F\}$, $Y=\{H,E\}$, $E=\{A\}$



With d-sep, we can know $P(X \mid E)\cdot P(Y \mid E) = P(X,Y \mid E)$

$\therefore P(F,H,E \mid A) = P(F \mid A)\cdot P(H,E \mid A)$

$\qquad \| \qquad\qquad\qquad\qquad \|$

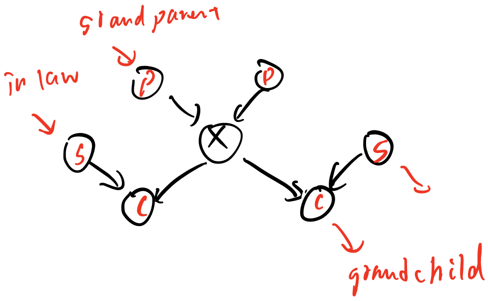
$P(F,H \mid E,A)\cdot P(E \mid A) \qquad P(H \mid A,E)\cdot P(E \mid A)$

$\Rightarrow P(F,H \mid E,A) = P(F \mid A)\cdot P(H \mid A,E)$.

13. Markov Blanket: (B_x)

$B_x := \{$ nodes consist of parent(x), children(x), spouses(x) $\}$

Thm: $P(x \mid B_x, Y) = P(x \mid B_x)$, if $Y \notin \{X, B_x\}$



14. Enumeration:

add & # of multiplications of $P(B=1 \mid J=1, M=1)$

$P(B,J,M) = \sum_{i,j} P(B=1, J=1, M=1, E=e_i, A=a_j) = \sum_{i,j} P(B=1)\cdot P(E=e)\cdot P(A=1 \mid B=1, E=e)\cdot P(J=1 \mid A=a) \cdot P(M=1 \mid A=a)$

16 multiplications, 3 additions

$$P(B \mid J,M) = \frac{P(B,J,M)}{P(B,J,M) + P(\bar{B},J,M)}$$

↗ redundant

$P(B,J,M)$ ↓ $16*$ $3+$

$P(\bar{B},J,M)$ ↓ $16*$ $3+$

$\therefore$ total: $32*$, $7+$.

query ↗ Evidence

Day 6 Variable Elimination: $P(Q \mid E)$

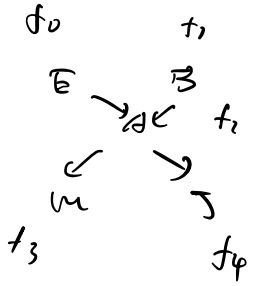
- Factor: a table that describes a function maps some values to vars (not necessarily just for prob)

V_1 V_2 ...	Values
	⋮

- Operations on Factors: { Multiplication, Addition, conditioning }

- eg: $p(B \mid J=1, M=1) = \dfrac{P(B, J=1, M=1)}{P(J=1, M=1)}$

$$= \frac{\Sigma_e \Sigma_a P(B, J=1, M=1, E=e, A=a)}{\cdots + \cdots}$$



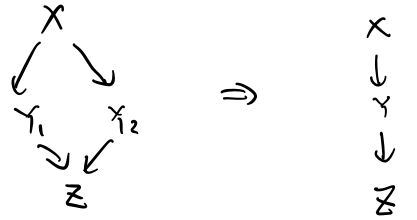
$$P(B, J=1, M=1, E=e, A=a) = P(E=e) \cdot P(B) \cdot P(A=a \mid B, E=e) \cdot P(J=1 \mid A=a) \cdot P(M=1 \mid A=a)$$

$$= f_0(e) \cdot f_1(B) \cdot f_2(a \mid B, e) \, f_3(J=1, A) \, f_4(M=1, a)$$

- Order of Elimination: (non-query, non-evidence) leaves → root, and head → bottom.

- polytree: A graph with no undirected loops.

collapsing nodes:



- How to elimate varibles:

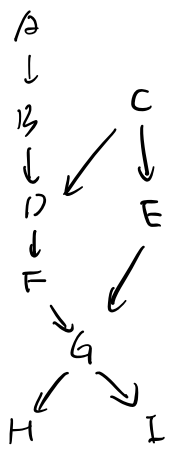
① create factors for CPTs

② elimating order (leaves → parents)

③ operate { multiple, summing out → $\Sigma_e \Sigma_a P(B, J=1, M=1, A=a, E=e)$ (marginal), conditioning

↳ $P(B, J=1)$

eg:



each variable is binary

we have all CPTs

Calculate $P(G|H=1)$

I. # of ops we need to do if we just use joint prob

$P(G|H=1) \propto \begin{cases} P(G=0|H=1) \\ P(G=1|H=1) \end{cases} \leftarrow P(G=0, H=1) \cdot P(G=1, H=1)$

$$P(G=0, H=1) = \sum_{a,b,c,d,e,f,i} P(G=0, H=1, A=a, B=b, C=c, D=d, E=e, F=f, I=i)$$

$$\overset{\text{CPTs}}{=} \sum_{abcdefi} P(A=a) \cdot P(B=b|A=a) \cdot P(C=c) \cdot P(D=d|B=b, C=c) \cdot P(E=e|C=c) \cdot P(F=f|D=d) \cdot P(G=0|F=f, E=e) \cdot P(H=1|G=0) \cdot P(I=i|G=0)$$

$\therefore 2^7$ terms, each term has 8*

# of multiplication $= 2 \times (2^7 \cdot 2^3) = 2^{11}$

# of addition $= 2 \times (2^7 - 1) = 2^8 - 2$

II. Calculate $P(G|H=1)$

$$P(G, H=1) = \sum_{a,b,c,d,e,f,i} P(G, H=1, A=a, B=b, C=c, D=d, E=e, F=f, I=i)$$

$$\overset{\text{CPTs}}{=} \sum_{abcdefi} P(A=a) \cdot P(B=b|A=a) \cdot P(C=c) \cdot P(D=d|B=b, C=c) \cdot P(E=e|C=c) \cdot P(F=f|D=d) \cdot P(G|F=f, E=e) \cdot P(H=1|G) \cdot P(I=i|G)$$

(Elimination Order: I, H, A, B, C, D, E, F)

① I, H, A, B:

$$\sum_{B=b} \left(P(D=d|B=b, C=c) \sum_{A=a} P(A=a) \cdot P(B=b|A=a) \right) \cdot P(H=1|G) \sum_{I=i} P(I=i|G)$$

(4*, 8*, 2*, 4*, condition, 2*)

$P(B, A)$

$F(B)$

$F(C, D)$

C	D	F(C, D)
0	0	...
...	...	...

$P(H=1|G)$ $\sum_{I=i} P(I=i|G)$ → $F(G)$

2*

G	F(H=1, G)
1	...
0	...

③ C. D. E. F :

$$\sum_{e,f} P(G \mid F=f, E=e) \cdot \sum_{D=d} P(F=f \mid D=d) \sum_{C=c} P(C=c) \cdot P(E=e \mid C=c) \cdot F(C,D) \cdot F(G)$$

2× 6+ 8× 4+ 8× 4+ 8×

$$= F(G) \cdot \sum_{e,f} F(G,F,E) \sum_{d} F(F,D) \sum_{c} F(C)\, F(E,C) \cdot F(C,D)$$

$F(C,D,E)$

$F(D,F)$

$F(G,F,E) \leftarrow F(F,E)$

$F'(G)$

∴ 40×, 20+ (decrease from 2^{11}×, 2^8+)

· Does Elimination order matter? Yes.

I. Time. Space of VE is dominated by the largest factor

II. Elimate variables that lead to the smallest next factor.

III. In polytree : linear time inference

Day 7. Learning Bayes Net. { learning parameters : CPTs ; learning structure : NP-hard.

· Problem : Bayes Net includes DAG / CPTs.

but sometimes we don't know them fully. we may need to learn the DAG or CPTs

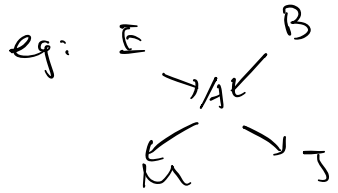
- In this leeture. we assume a given DAG and learn the CPTs from data :

  - with complete data : MLE
  - with incomplete data : EM

- ML (Maximum likelyhood) learning:

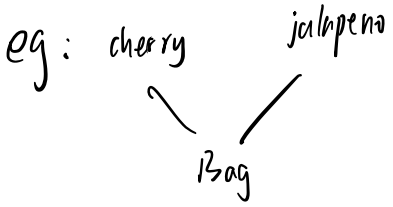
  - Idea: Choose the model (CPTs) that maximizes the Prob of data.

    i.e. $P_{model}(\text{data})$

eg: 

Sample	E	B	A	J	M
1	0	0	1	1	1
2	1	0	1	0	0
3					
⋮					

that is: $\underset{CPTs}{\text{argmax}}\ P(\text{data} \mid CPTs)$

eg: 

DAG: (X)

$X: \{\text{cherry}, \text{jalapeno}\}$

$CPT: P(X = \text{cherry}) = p$

$\text{data}: \{x^{(1)}, x^{(2)} \dots x^{(T)}\}$ — T samples.

$$\begin{aligned}
&\underset{p}{\text{argmax}}\ P(\text{data} \mid p) \\
&= \underset{p}{\text{argmax}}\ \underbrace{P(x^1 \cdots x^T \mid p)}_{\text{likelyhood}} \\
&= \underset{p}{\text{argmax}} \prod_{i=1}^{T} P(x^{(i)} \mid p) \\
&= \underset{p}{\text{argmax}} \log \prod_{i=1}^{T} P(x^{(i)} \mid p) \\
&= \underset{p}{\text{argmax}} \log p^{N_c}(1-p)^{N_j} \\
&= \underset{p}{\text{argmax}} \left(N_c \cdot \log p + N_j (1-p)\right)
\end{aligned}$$

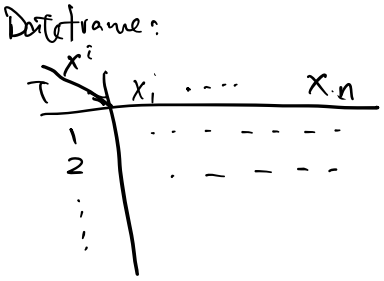
$\therefore$ log-likelyhood: $L(p) = N_c \log p + N_j (1-p)$

$\therefore \frac{\partial L(p)}{\partial p} = 0 \iff p = \frac{N_c}{N_c + N_j} = \frac{N_c}{T}$

Estimating para for B.N.: $p(X_i = x \mid Pa(X_i) = \pi)$

Data: $\{(x_1^{(t)}, \cdots x_n^{(t)})\}_{t=1}^{T}$

Dataframe:



$$\ell = \log P(\text{data}) \overset{i.i.d}{=} \log \prod_{t=1}^{T} p(X_1 = x_1^t, \cdots X_n = x_n^t)$$

$$P(X_1^{(t)}, \cdots X_n^{(t)}) = \prod_{i=1}^{n} P(X_i = x_i^{(t)} \mid Pa(X_i) = Pa\, x_i^{(t)})$$

$$\Rightarrow \ell = \log \prod_{t=1}^{T} \prod_{i=1}^{n} P(X_i = x_i^{(t)} \mid Pa(X_i) = Pa(x_i^{(t)})$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n} \log P(X_i = x_i^{(t)} \mid Pa(X_i) = Pa(x_i^{(t)}))$$

(sum all samples; sum all variables)

(for single varible: $L = \sum_{t=1}^{T} \log(P(X = x^{(t)})$

- def: $\text{count}(X_i = x_i^{(t)}, Pa(X_i) = Pa(x_i^{(t)})) \overset{\text{def}}{=}$ # of samples where $X_i = x_i^{(t)}, Pa(x_i) = Pa(x_i^t)$

$$\Rightarrow \ell = \sum_{i=1}^{n} \sum_{t=1}^{T} \log P(x_i^{(t)} \mid Pa_i^{(t)})$$

$$= \sum_{i=1}^{n} \sum_{x} \sum_{\pi} \underbrace{\text{count}(X_i = x, Pa_i = \pi)}_{\text{constant}} \log P(X_i = x \mid Pa_i = \pi)$$

we need: $\dfrac{\partial \ell}{\partial P(X_i = x \mid Pa_i = \pi)} = 0$

$$\Rightarrow P_{ML}(X_i = x \mid Pa_i = \pi) = \frac{\text{count}(X_i = x, Pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x', Pa_i = \pi)}$$

$$\text{or } P_{ML}(X_i = x) = \frac{\text{count}(X_i = x)}{T}$$

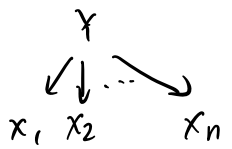
def: $I(x, y) \overset{\text{def}}{=} \begin{cases} 1, & x = y \\ 0, & \text{otherwise.} \end{cases}$

then we know: $P_{ML}(X_i = x) = \frac{1}{T}\text{count}(X_i = x) = \frac{1}{T}\sum_{t=1}^{T} I(x_i^{(t)}, x)$

eg: $P(F = f \mid E = e, D = d)$

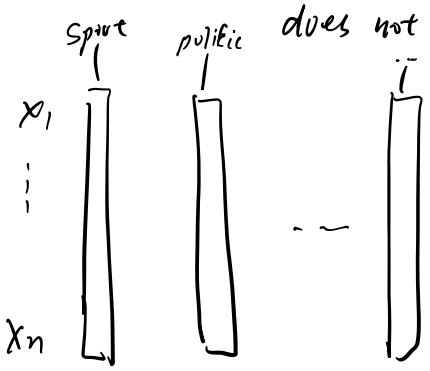
$$= \frac{\sum_t I(f_t, f)\, I(e_t, e)\, I(d_t, d)}{\sum_t I(c_t, c)\, I(d_t, d)}$$

✡ Naive Bayes: a Bayes Net with a single parent and many children.



eg: $Y \in \{1, 2, \dots k\}$, 1=sports, 2=fashion, 3=...

$x_i \in \{0, 1\}$, for $i = 1 \dots n$, where 0 means ith word in dictionary does not appear in doc



learning:

- goal: observe a set of data $\{x_1, \dots x_n\}$, try to inference Y.

- from data: $P(x_i = x \mid Y = y) = \dfrac{\sum_t I(y_t, y) \cdot I(x_i^t, x)}{\sum_t I(y_t, y)}$

Classification:

$$P(Y = y \mid \{X_1 = x_1, \dots, X_n = x_n\}) = \frac{P(Y=y) \prod_{i=1}^n P(X_i = x_i \mid Y = y)}{\sum_{y'} P(Y = y') \prod_{i=1}^n P(X_i = x_i \mid Y = y')}$$

✡ likelyhood and KL distance:

log-likelyhood: $L(p) = \log P_{model}(data)$

$$KL(q, p) = \sum_{k=1}^n q_k \log \frac{q_k}{p_k}$$

Situation: An n-sided die is tossed T times. data: $\{x^{(1)}, \dots x^{(T)}\}$.

In the course of T tosses: $C_k \overset{def}{=}$ # of the k^{th} side observed.

so: $L(p) = \log P_{model}(data) = \log p(\{x^{(1)}, \dots, x^{(T)}\}) = \log \prod_{t=1}^T P(X = x^{(t)}) = \sum_{t=1}^T \log P(X = x^{(t)})$

$$= \sum_{k=1}^n C_k \log P_k$$

define $q_k = \frac{C_k}{T}$ as a distribution:

$$KL(q, p) = \sum_{k=1}^n q_k \log q_k - \sum_{k=1}^n q_k \log p_k$$

$$= \sum_{k=1}^n q_k \log q_k - \frac{\sum_{k=1}^n C_k \log p_k}{T}$$

$$= \sum_{k=1}^n q_k \log q_k - \frac{L(p)}{T}$$

∴ conclusion:

maximizing $L(p)$ $\Leftrightarrow$ minimizing $KL(q, p)$

Day 9: EM Algorithm (Expectation Algorithm):

- goal: estimate all the CPTs.

- from MLE (all data observed):

result:
$$\begin{cases} P(X=x_i) = \dfrac{count(X=x_i)}{T} \\ P(X=x_i \mid Pa_i=\pi) = \dfrac{count(X=x_i, Pa_i=\pi)}{count(Pa_i=\pi)} = \dfrac{\sum_t I(X_i^{(t)}, x_i)\, I(Pa_i^{(t)}, \pi)}{\sum_t I(Pa_i^{(t)}, \pi)} \end{cases}$$

- learning with partial data, generally:

Nodes: $\{X_1, \dots X_n\}$ in BN.

unobserved nodes: H

observed nodes: V

$H \cup V = \{X_1, \dots, X_n\}$. $H \cap V = \phi$

→ goal: estimate CPTs to maximize Prob of given data.

i.e. $\underset{CTs}{\operatorname{argmax}}\ p(\text{data} \mid CPTs)$

→ method: $\underset{CPTs}{\operatorname{argmax}}\ \log L$, $L = \log \prod_{t=1}^{T} P(V=v^{(t)}) = \sum_{t=1}^{T} \log P(V=v^t)$

$$= \sum_{t=1}^{T} \log \sum_{h} P(V=v^t, H=h)$$

$$= \sum_{t=1}^{T} \log \sum_{h} \left(\prod_{i=1}^{n} P(X_i=x_i \mid Pa_i=\pi_i) \right)$$

$$\frac{\partial L}{\partial P(X_i=x_i \mid Pa_i=\pi)} = 0 \quad \Rightarrow \text{no closed form solution.}$$

→ Idea: we use a iterative algorithm.

✱ EM (Expectation Maximization) alg:

- Node with parents: $P(X_i = x \mid Pa_i = \pi) = \frac{count(X_i = x, Pa_i = \pi)}{count(Pa_i = \pi)}$
- root nodes: $P(X_i = x) = \frac{count(X_i = x)}{T}$

- EM alg:

  Init para of the model

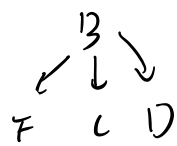
  while para not converged:

  For each para p:

  - E-step: Calculate expected counts by using para from last iteration
  - M-step: use expected counts to update p

- Initialization:
  - random guess.
  - in $[0,1]$
  - EM may not produce global max

eg:



(unobserved)

(observed)

we want $\begin{cases} P(B) \\ P(F|B),\ P(C|B),\ P(D|B) \end{cases}$

$P(B=1) = \frac{count(B=1)}{T}$ → we don't have.

→ use expected count $= \sum_{t=1}^{T} P(B=1 \mid F=f^{(t)}, C=c^{(t)}, D=d^{(t)})$

$P(B=1 \mid F=f^{(t)}, C=c^{(t)}, D=d^{(t)})$

$= \frac{P(B=1, F=f^{(t)}, C=c^{(t)}, D=d^{(t)})}{P(B=1, F=f^{(t)}, C=c^{(t)}, D=d^{(t)}) + P(B=0, F=f^{(t)}, C=c^{(t)}, D=d^{(t)})}$

$P(B=1, F=f^{(t)}, C=c^{(t)}, D=d^{(t)}) = P(B) \cdot P(F=f^{(t)} \mid B) \cdot P(C=c^{(t)} \mid B) \cdot P(D=d^{(t)} \mid B)$

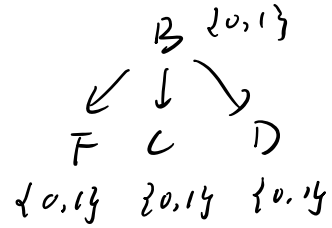
previous guess

$$\text{count}(B=1) = \sum_{t=1}^{T} P(B=1 \mid F=f^{(t)}, C=c^{(t)}, D=d^{(t)})$$

$$= \sum_{t=1}^{T} \frac{P(B=1)\cdot P(F=f^{(t)} \mid B=b)\cdot P(C=c^{(t)} \mid B=b)\cdot P(D=d^{(t)} \mid B=1)}{\sum_b P(B=b)\cdot P(F=f^{(t)} \mid B=b)\cdot P(C=c^{(t)} \mid B=b)\cdot P(D=d^{(t)} \mid B=b)}$$

E-step : use above method to calculate count(B=1)

M-step: update $P(B=1) \leftarrow \frac{\text{count}(B=1)}{T}$



$$P(F=1 \mid B=1) = \frac{\text{count}(B=1, F=1)}{\text{count}(B=1)}$$ ↳ done

$$\text{Count}(B=1, F=1) = \sum_{t=1}^{T} P(B=1, F=1 \mid F=f^{(t)}, C=c^{(t)}, D=d^{(t)})$$

$$= \sum_{t=1}^{T} P(B=1 \mid F=1, C=c^{(t)}, D=d^{(t)}) \cdot I(f^{(t)}, 1)$$

$$= \sum_{t=1}^{T} \frac{P(B=1, F=1, C=c^{(t)}, D=d^{(t)}) \cdot I(f^{(t)}, 1)}{\sum_b P(B=b, F=1, C=c^{(t)}, D=d^{(t)})}$$

$$= \sum_{t=1}^{T} \frac{P(B=1)\cdot P(F=1 \mid B=1)\cdot P(D=d^{(t)} \mid B=1)\cdot P(C=c^{(t)} \mid B=1)\cdot I(f^{(t)}, 1)}{\sum_b P(B=b)\cdot P(F=1 \mid B=b)\cdot P(D=d^{(t)} \mid B=b)\cdot P(C=c^{(t)} \mid B=b)}$$

↳ from last iteration.

Generally, EM:

H — hidden nodes

V — visible nodes

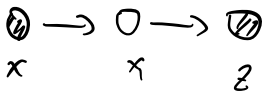
Nodes with parents: $P(X_i=x \mid Pa_i=\pi) = \frac{\text{count}(X_i=x, Pa_i=\pi)}{\text{count}(Pa_i=\pi)}$

$$= \frac{\sum_t P(X_i=x, Pa_i=\pi \mid \vec{V}^{(t)})}{\sum_t P(Pa_i=\pi \mid \vec{V}^{(t)})}$$

roots: $P(X_i=x) = \frac{\text{count}(X_i=x)}{T} = \frac{\sum_t P(X_i=x \mid \vec{V}^{(t)})}{T}$

• EM properties:

1. Monotonic convergence: each iteration increases log-likelyhood of observed data
2. No tunning para, lr. backtracking.
3. converges to a local or global maximum.

eg: 

use EM alg to update $P(Z=z \mid Y=y)$

$V=\{X, Z\}$, $H=\{\}$

$$P(Z=z \mid Y=y) = \frac{\text{count}(Z=z, Y=y)}{\text{count}(Y=y)}$$

$$= \frac{\sum_t P(Z=z, Y=y \mid X=x^{(t)}, Z=z^{(t)})}{\sum_t P(Y=y \mid X=x^{(t)}, Z=z^{(t)})}$$

$$P(Y=y \mid X=x^{(t)}, Z=z^{(t)})$$

$$= \frac{P(Y=y, X=x^{(t)}, Z=z^{(t)})}{\sum_{y'} P(Y=y', X=x^{(t)}, Z=z^{(t)})}$$

$$= \frac{\sum_t P(Y=y \mid X=x^{(t)}, Z=z^{(t)}) \cdot \mathbb{I}(Z, z^{(t)})}{\sum_t P(Y=y \mid X=x^{(t)}, Z=z^{(t)})}$$

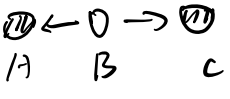
$$P(Y=y, X=x^{(t)}, Z=z^{(t)})$$

$$= P(X=x^{(t)}) \cdot P(Y=y \mid X=x^{(t)}) \cdot P(Z=z^{(t)} \mid Y=y)$$

$$\therefore P(Z=z \mid Y=y) = \frac{\sum_t \frac{P(X=x^{(t)}) \cdot P(Y=y \mid X=x^{(t)}) \cdot P(Z=z^{(t)} \mid Y=y)}{\sum_{y'} P(X=x^{(t)}) \cdot P(Y=y \mid X=x^{(t)}) \cdot P(Z=z^{(t)} \mid Y=y)} \cdot \mathbb{I}(Z, z^{(t)})}{\sum_t \frac{P(X=x^{(t)}) \cdot P(Y=y \mid X=x^{(t)}) \cdot P(Z=z^{(t)} \mid Y=y)}{\sum_{y'} P(X=x^{(t)}) \cdot P(Y=y \mid X=x^{(t)}) \cdot P(Z=z^{(t)} \mid Y=y)}}$$

↳ CPTs from last iteration.

(binary variables)

eg2: 

A, C observed

B unobserved

Data: $\{(a_t, c_t)\}_{t=1}^{T}$

	a_t	c_t
1	0	1
2		
⋮		
T	1	1

CPTs: $P(B=1)$, $P(A=1 \mid B=1)$, $P(C=1 \mid B=1)$, $P(A=1 \mid B=0)$, $P(C=1 \mid B=0)$

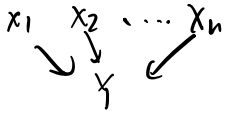
$$P(B=1) = \frac{\text{count}(B=1)}{T} = \frac{\sum_t P(B=1 \mid A=a^{(t)}, C=c^{(t)})}{T}$$

$$P(A=1 \mid B=1) = \frac{\text{count}(A=1, B=1)}{\text{count}(B=1)} = \frac{\sum_t P(A=1, B=1 \mid A=a^{(t)}, C=c^{(t)})}{\sum_t P(B=1 \mid A=a^{(t)}, C=c^{(t)})}$$

$$\text{CPTs} \leftarrow \text{Full Joint} = \frac{\sum_t P(B=1 \mid A=a^{(t)}, C=c^{(t)}) \cdot \mathbb{I}(A, a^{(t)})}{\sum_t P(B=1 \mid A=a^{(t)}, C=c^{(t)})}$$

other CPTs can be updated in the same way.

※ EM for learning Noisy-OR model.



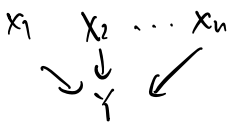
- Data: $\{(x_1^{(t)}, x_2^{(t)}, \cdots, x_n^{(t)}, y^{(t)})\}_{t=1}^{T}$
- $P(Y=1 \mid X_1, X_2 \cdots X_n) = 1 - \prod_{i=1}^{n} (1-p_i)^{x_i}$

- problem: estimate $p_i \in [0,1]$

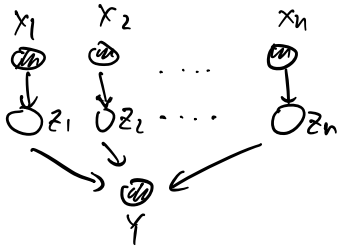
- Why don't we just use MLE since we have fully observed data?

  → we can't get a closed form.

- Method:



Transformed



CPTs:

$P(z_i=0 \mid x_i=0) = 1$

$P(z_i=1 \mid x_i=1) = p_i$

$P(Y \mid \vec{z}) = OR(\vec{z})$

- First show that: $P(Y=1 \mid \vec{x}) = \sum_{\vec{z}} P(Y=1 \mid \vec{z}) \cdot P(\vec{z} \mid \vec{x})$

$$P(Y=1 \mid \vec{x}) = \sum_{\vec{z}} P(Y=1, \vec{z} \mid \vec{x}) = \sum_{\vec{z}} P(Y=1 \mid \vec{x}, \vec{z}) \cdot P(\vec{z} \mid \vec{x})$$

$$= \sum_{\vec{z}} P(Y=1 \mid \vec{z}) \cdot P(\vec{z} \mid \vec{x})$$

- Q: when does $P(Y=1 \mid \vec{z}) \cdot P(\vec{z} \mid \vec{x}) = 0$?

  → all $z_i \in \vec{z}$ equals to 0.

- $P(Y=1 \mid \vec{x}) = \sum_{\vec{z}} P(Y=1 \mid \vec{z}) \cdot P(\vec{z} \mid \vec{x}) = \sum_{\vec{z} \neq 0} P(\vec{z} \mid \vec{x}) = 1 - \sum_{\vec{z}=0} P(\vec{z} \mid \vec{x})$

- update $P(z_i=1 \mid x_i=1)$ by EM alg:

$$P(z_i=1 \mid x_i=1) = \frac{\text{Count}(z_i=1, x_i=1)}{\text{Count}(x_i=1)} \longrightarrow \text{from data}$$

Data:

x_1	....	x_n	Y	t
0	....	1	1	1
⋮		⋮	0	2
⋮		⋮	⋮	?
⋮		⋮	1	⋮
0		1	0	T

$$\text{count}(z_i=1, x_i=1) = \sum_{t=1}^{T} P(z_i=1, x_i=1 \mid \vec{x}^{(t)}, y^{(t)})$$

$$= \sum_{t=1}^{T} x_i^{(t)} \cdot P(z_i=1 \mid \vec{x}^{(t)}, y^{(t)})$$

($x_i^{(t)} \Leftrightarrow I(x_i^{(t)}, 1)$, numerically)

$$P(z_i=1 \mid \vec{x}^{(t)}, y^{(t)}) = \frac{P(Y=y^{(t)} \mid z_i=1) \cdot P(z_i=1 \mid \vec{x}^{(t)})}{P(Y=y^{(t)} \mid \vec{x}^{(t)})}$$

$$= \frac{y^{(t)} \cdot P_i}{1 - \prod_{i=1}^{n} (1-P_i)^{x_i}}$$

$y^{(t)}$: OR relationship

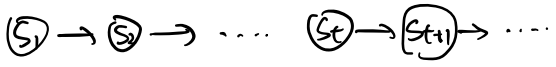
$P_i = P(z_i=1 \mid \vec{x})$; $x_i=1$ and d-sep

$$\Rightarrow \text{count}(z_i=1, x_i=1) = \sum_{t=1}^{T} \left[\frac{P_i \cdot y^{(t)} \cdot x_i^{(t)}}{1 - \prod_{j=1}^{n} (1-P_i)^{x_i}} \right]$$

$$\Rightarrow P_i \leftarrow \frac{\text{Count}(z_i=1, x_i=1)}{\text{Count}(x_i=1)} = \frac{P_i}{T_i} \sum_{t=1}^{T} \left[\frac{y^{(t)} \cdot x_i^{(t)}}{1 - \prod_{j=1}^{n} (1-P_j)^{x_j}} \right]$$

Day 10 HMM (Hiden markov model):

- Markov chain:



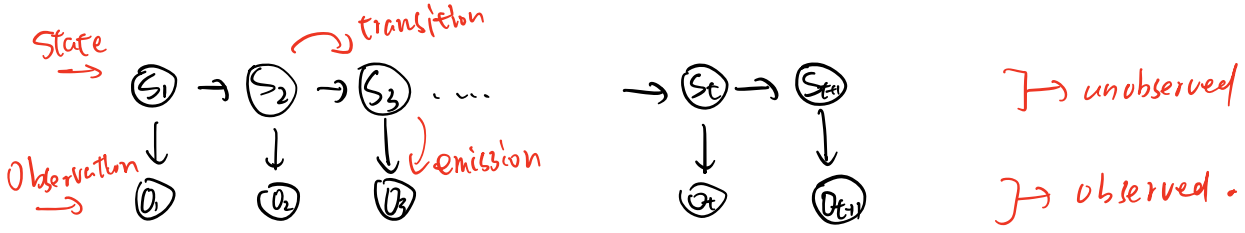
2 assumptions:

I. $P(S_t \mid S_1 \cdots S_{t-1}) = P(S_t \mid S_{t-1})$ —— finite context

II. $P(S_{t+1}=j \mid S_t=i) = P(S_{t+k+1}=j \mid S_{t+k}=i)$ —— position invariant.

(we can save lots of CPTs with these 2 assumptions)

- HMM: extension of Markov chain.



Assumptions:

I. $P(S_t \mid S_1 \cdots S_{t-1}) = P(S_t \mid S_{t-1})$

$P(O_t \mid S_1 \cdots S_t) = P(O_t \mid S_t)$

II. $P(S_{t+1} \mid S_t) = P(S_{t+k+1} \mid S_{t+k})$

$P(O_t \mid S_t) = P(O_{t+k} \mid S_{t+k})$

(CPTs are matrices)

eg: HMM for language:

States: AI → is → fun

Observation: 

(sound waves)

★ Conditional Independence in HMM:

- $P(S_t|S_{t-1}) \neq P(S_t|S_{t-1}, O_t)$ is obvious, which means observation affects the prob.
- $P(S_2, S_3 \cdots S_T|S_1) = \prod_{t=2}^{T} P(S_t|S_{t-1})$

Show it:

$$P(S_2 \cdots S_T|S_1) = P(S_T|S_{T-1} \cdots S_1) \cdot P(S_2 \sim S_{T-1}|S_1)$$

$$= P(S_T|S_{T-1} \cdots S_1) \cdot P(S_{T-1}|S_1 \cdots S_{T-2}) \cdot P(S_{T-2} \cdots S_2|S_1)$$

$$\vdots$$

$$= P(S_T|S_{T-1} \cdots S_1) \cdot P(S_{T-1}|S_{T-2} \cdots S_1) \cdots P(S_2|S_1)$$

$$= P(S_T|S_{T-1}) \cdots P(S_2|S_1)$$

$$= \prod_{t=2}^{T} P(S_t|S_{t-1})$$

- Compute: $P(O_1, O_2 \cdots O_T)$

$$P(O_1, O_2 \cdots O_T) = \sum_{i=1}^{n} P(O_1, \cdots O_T, S_T = n)$$

Def: $\alpha_t(i) = P(O_1 \cdots O_t, S_t = i)$

transition prob : $a_{ij} = P(S_t = i | S_{t-1} = j)$

emission prob : $b_{jO_k} = P(O_t = O_k | S_t = j)$

init prob : $\pi_i = P(S_1 = i)$

We want to get a recurent relationship of α_{it}

$$\alpha_T(S_T) = P(O_1, \cdots O_T, S_T)$$

$$= \sum_{k=1}^{n} P(O_1 \cdots O_T, S_T, S_{T-1} = k)$$

$$= \sum_{k=1}^{n} P(O_T | S_T, S_{T-1} = k, O_1 \cdots O_{T-1}) \cdot P(S_T | S_{T-1} = k, O_1 \cdots O_{T-1}) \cdot P(O_1, \cdots, O_{T-1}, S_{T-1} = k)$$

$$\overset{d\text{-sep}}{=} \sum_{k=1}^{n} p(O_T \mid S_T) \cdot p(S_T \mid S_{T-1}=k) \cdot p(O_1 \cdots O_{T-1}, S_{T-1}=k)$$

$$= \sum_{k=1}^{n} p(O_T \mid S_T) \cdot p(S_T \mid S_{T-1}=k) \cdot \alpha_{T-1}(S_{T-1}=k)$$

emission　　transition　　recursion.

$$\therefore \alpha_{j,t+1} = \sum_{i=1}^{n} \alpha_{it} a_{ij} b_j o_{t+1} \quad , \alpha_{i1} = \pi_i b_{i o_1} \longrightarrow p(O_1, S_1) = p(O_1 \mid S_1) \cdot p(S_1)$$

$$\therefore p(O_1, \cdots, O_T) = \sum_{i=1}^{n} \alpha_T(i) = \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_{T-1}(k)\, a_{ij}\, b_j O_T$$

for N states. N observation, complexity: $\Theta(n^2 N)$