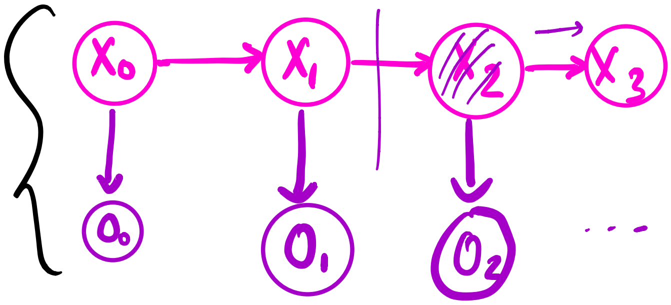
HMM.

Recap. what is a HMM?

$$\lambda = \{ \underset{S}{\overset{N}{\{s_1 \dots s_N\}}}, \underset{\Theta}{\overset{M}{\{\theta_1 \dots \theta_M\}}}, a_{ij}, b_i(\ell), \pi \}$$



$x_t \in S, \; O_t \in \Theta$

$\pi_i = Pr(X_0 = i)$

→ $a_{ij} = Pr(X_{t+1} = s_j \mid X_t = s_i)$, $b_i(\ell) = Pr(O_t = \theta_\ell \mid X_t = s_i)$

'T' → total time.

1. State prob. estimation. (Forward-Backward Alg)

Estimate $Prob(X_t = s_j \mid O_1 \dots O_T, \lambda)$.
(not t)

2. Most Probable Path (MPP). (viterbi Algorithm)

Given $O_1 \dots O_T$, and λ

$$\max_{s_0 s_1 \dots s_T} Prob(X_0 = s_0, X_1 = s_1 \dots X_T = s_T \mid O_1 \dots O_T, \lambda)$$

3. Learn the HMM (Baum Welch Algorithm)

→ EM algorithm

* $$\hat{\lambda} = \arg\max_{\lambda} Pr(O_1 \dots O_T \mid \lambda).$$

$\hat{\lambda} = \{ \hat{a}_{ij}, \hat{b}_i(\ell), \hat{\pi}, \{s_1 \dots s_N\}, \{\theta_1 \dots \theta_M\} \}$

- For 1, 2, we assume knowledge of the model.
-

1. State prob. estimation. (F B algorithm)

Estimate $\text{Prob}(X_{t+1} = s_j \mid O_0 \cdots O_T, \lambda)$

(Last time we did $\text{Prob}(X_{t+1} = s_j \mid O_0 \cdots O_{t+1}, \lambda)$, which is half of the puzzle.

Forward Alg.

- $\alpha_t(i) := \text{Prob}(O_0, \dots, O_t, X_t = s_i \mid \lambda) \longrightarrow$ (probability of past & ending in state i.

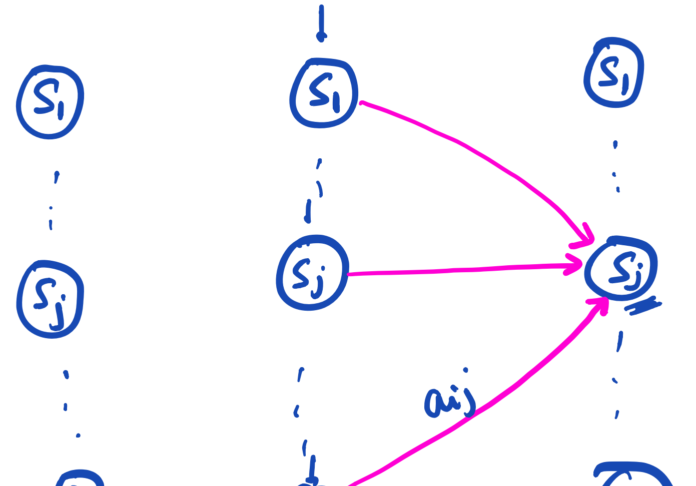
$T \times N$ $\alpha_t(j)$

$$\boxed{\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i) \cdot a_{ij}\right) b_j(O_{t+1})} \quad \text{— (I)}$$

$$\rightarrow \Pr(O_0 \cdots O_t \mid \lambda) = \sum_{i=1}^{n} \alpha_t(i) \quad \text{— (II)}$$

$\alpha_0(i) = \pi_i \cdot b_i(O_0)$
$= \text{Prob}(O_0, X_0 = s_i \mid \lambda)$

$N^2 T$



$$* \; \Pr(O_0 \cdots O_T \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

SN SN SN

$t \longrightarrow$ $t-1$ $\quad$ t $\quad$ $t+1$

O_{t-1} $\quad$ O_t $\quad$ O_{t+1}

Backward Algorithm

$$\beta_t(i) = Pr(O_{t+1}, \dots, O_T \mid x_t = s_i, \lambda).$$

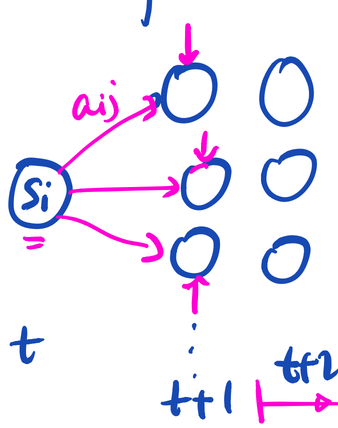
- Given state at time t, what is prob of observations in future.

$$\beta_t(i) = Pr(O_{t+1}, \dots, O_T \mid x_t = s_i, \lambda).$$

$$= \sum_{j=1}^{N} Pr(O_{t+1} \cdots O_T, \underbrace{x_{t+1} = j} \mid x_t = s_i, \lambda).$$

$$= \sum_{j=1}^{N} Pr(O_{t+2} \cdots O_T \mid x_{t+1} = j, \lambda)\, Pr(x_{t+1} = s_j \mid x_t = s_i, \lambda)\, \boxed{Pr(O_{t+1} \mid x_{t+1} = j, \lambda)}$$

$$\boxed{\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) \cdot a_{ij} \cdot b_j(O_{t+1})}$$



Let $\gamma_t(i) := Pr(x_t = s_i \mid O_0 \dots O_T, \lambda)$?

$$\underbrace{O_0 \cdots O_t}_{\alpha} \mid \underbrace{\cdots O_T}_{\beta}$$

$$\gamma_t(i) = \frac{Pr(x_t = s_i, O_0 \cdots O_T \mid \lambda)}{Pr(O_0 \cdots O_T \mid \lambda)}$$

$$= \frac{\overbrace{Pr(O_0 \dots O_t, X_t = S_i \mid \lambda)} \cdot \overbrace{Pr(O_{t+1} \cdots O_T \mid \lambda, X_t = S_i, \cancel{O_0 \dots O_t})}}{Pr(O_0 \dots O_T \mid \lambda)}$$

$P(A, B \mid \lambda)$
$= P(A \mid \lambda) \cdot P(B \mid \lambda, A)$

$$\boxed{\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}} \Rightarrow \boxed{\sum \gamma_t(i) = 1}$$

Prove $\quad Pr(O_0 \dots O_T \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$

$$= \sum_{i=1}^{N} Pr(O_0 \dots O_T, X_t = S_i \mid \lambda) \dots$$

2. * Most Probable Path (MPP). (viterbi Algorithm)

Given $O_1 \dots O_T$, and λ

$$\max_{S_0 S_1 \dots S_T} Prob(X_0 = S_0, X_1 = S_1 \dots X_T = S_T \mid \underline{\underline{O_1 \dots O_T}}, \underline{\underline{\lambda}}) \longrightarrow (*)$$

$X_t = ? \quad \longrightarrow X_t^* = \arg\max_i \gamma_t(i)$

$$= \arg\max_i Pr(X_t = S_i \mid O_0 \dots O_T, \lambda)$$

$(X_1^*, X_2^*, \dots, X_T^*) \longrightarrow$ not solving $(*)$

'q' → 'e' → 'q' → 'd'

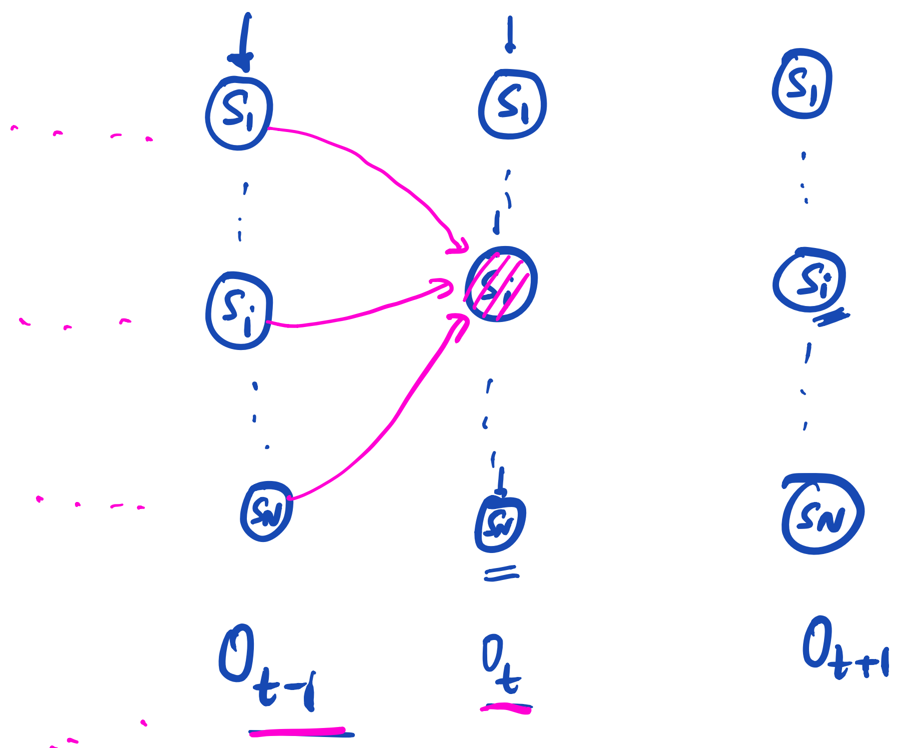
$$\ast \quad \delta_t(i) = \max_{x_1 \dots x_{t-1}} \Pr(x_0 \dots x_{t-1}, x_t = s_i, O_0 \dots O_t \mid \lambda) \quad \text{— ①}$$

$$\delta_{t+1}(j) = \max_{x_1 \dots x_t} \Pr(x_0 \dots x_t, x_{t+1} = s_j, O_0 \dots O_{t+1} \mid \lambda)$$

$$= \max_{i} \left\{ \max_{x_1 \dots x_{t-1}} \Pr(x_0 \dots x_{t-1}, x_t = s_i, x_{t+1} = s_j, O_0 \dots O_t, O_{t+1} \mid \lambda) \right\}$$

$$\vdots$$

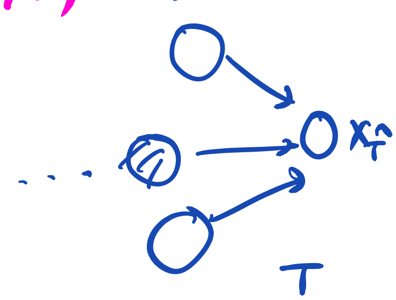
$$\boxed{\delta_{t+1}(j) = \max_{i=1 \dots N} \{ \delta_t(i) \cdot a_{ij} \} \, b_j(O_{t+1})} \quad \text{——} \ ☆$$



$$\delta_t(j) = \max_{i} \underbrace{\{ \delta_{t-1}(i) \cdot a_{ij} \}} \, b_j(O_t) \quad ✓$$

$$\max_j \delta_T(j) = \max_{x_0 \dots x_T} \Pr(x_0 \dots x_T, O_0 \dots O_T \mid \lambda)$$

$$\hat{x}_T = \arg\max_{j} [\delta_T(j)]$$



$$\hat{x}_{t-1} = \arg\max \left[\delta_{t-1}(j) \cdot a_{j \hat{x}_t} \right]$$

3. Learn the HMM (Baum Welch Algorithm) → EM algorithm

$$\boxed{\hat{\lambda} = \arg\max_{\lambda} \Pr(O_0 \dots O_T \mid \lambda)}$$

Problems 1 & 2 were inference tasks.

$$\lambda = \Big\{ \underbrace{\{S_1 \dots S_N\}}_{S},\ \underbrace{\{O_1 \dots O_M\}}_{\Theta},\ a_{ij},\ b_i(\ell),\ \pi \Big\}$$

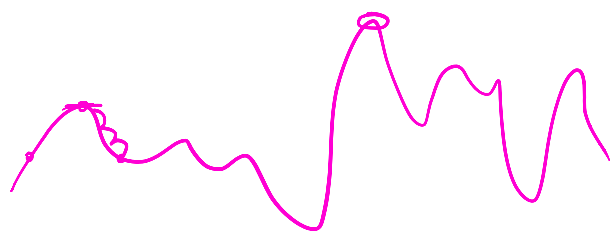
Given $\boxed{N, M}$ $\quad \Pr(O_0 \dots O_T \mid \lambda) \longrightarrow 1.$

'T+1' states $S_0\ S_1 \dots S_T$ $\quad X_0 = S_0, \dots\ X_T = S_T$, $\pi_0 = 1$,

$a_{t,t+1} = 1$, $\quad b_j(O_j) = 1.$

This optimization is HARD,

$$\mathbb{R}^{\{\underbrace{N^2}_{a_{ij}} + \underbrace{N \cdot M}_{b_j(\ell)} + \underbrace{N}_{\pi}\}}$$



① K- means algorithm

② recommendor systems.

EM algorithm

Dempster.... wiki.

- An iterative procedure.

$$\lambda^{(0)} = \{ \pi_i^{(0)}, a_{ij}^{(0)}, b_j^{(0)}(\ell) \ldots \} \longrightarrow \lambda^{(1)}, \ldots$$