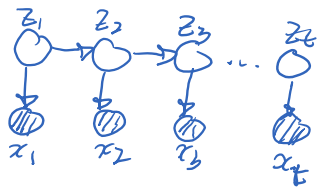


today: HMM inference & learning

HMM (hidden Markov model)



$z_t \in \{1, \dots, K\}$ discrete

(later $z_t \sim$ Gaussian)

x_t ds. e.g. speech signal

\leadsto Kalman filter

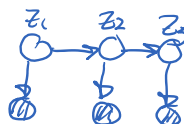
discrete e.g. DNA sequence

HMM \rightarrow generalization of mixture model

GMM



add time dependence



$$DGM: p(x_{1:T}, z_{1:T}) = p(z_1) \prod_{t=1}^T \underbrace{p(x_t | z_t)}_{\text{emission prob.}} \prod_{t=2}^T \underbrace{p(z_t | z_{t-1})}_{\text{transition prob.}}$$

often, emission & trans probs are homogeneous in time (i.e. don't depend on t)

$$p_t(x_t | z_t) = f(x_t | z_t) \quad [\text{pmf or pdf on } x_t]$$

$$p_t(z_t = i | z_{t-1} = j) = A_{ij}$$

"stochastic matrix"

$$A = \begin{pmatrix} & j \\ i & \end{pmatrix} \rightarrow \text{def. over } \mathbb{Z}$$

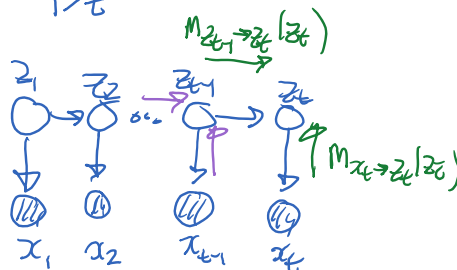
$$\sum_i A_{ij} = 1 \quad \forall j$$

inference tasks:

- prediction $p(z_t | x_{1:t-1})$ "where next"
- filtering $p(z_t | x_{1:t})$ "where now"
- smoothing $p(z_t | x_{1:T})$ "where in the past?" $T > t$

α -recursion:

let's run sum-product here to derive recursions to compute following prob



to compute $p(z_t, \bar{x}_{1:t})$ or $p(z_t | \bar{x}_{1:t})$

$$p(z_t, \bar{x}_{1:t}) = \underbrace{1}_{\alpha_t(z_t)} \cdot M_{z_{t-1} \rightarrow z_t}(z_t) \cdot m_{x_t \rightarrow z_t}(z_t) \quad (\text{have } z=1)$$

$$m_{x_t \rightarrow z_t}(z_t) = \sum_{\bar{x}_t} p(x_t | z_t) \delta(x_t, \bar{x}_t) \stackrel{\text{conditioning trick}}{=} p(\bar{x}_t | z_t)$$

$$\alpha_t(z_t) = \sum_{x_t} p(x_t|z_t) S(x_t, \bar{x}_t) = p(\bar{x}_t | z_t)$$

$$M_{z_{t-1} \rightarrow z_t}(z_t) = \sum_{z_{t-1}} p(z_t|z_{t-1}) M_{z_{t-2} \rightarrow z_{t-1}}(z_{t-1}) \cdot M_{x_{t-1} \rightarrow z_{t-1}}(z_{t-1})$$

$$\alpha_t(z_t) = p(\bar{x}_t | z_t) \sum_{z_{t-1}} p(z_t|z_{t-1}) \alpha_{t-1}(z_{t-1})$$

$\underbrace{p(\bar{x}_t | z_t)}_{\text{vector}(z_t)}$
 $\underbrace{p(z_t|z_{t-1})}_{\text{matrix}}$
 $\underbrace{\alpha_{t-1}(z_{t-1})}_{\text{vector}(z_{t-1})}$

α -recursion aka. "forward recursion" like the "collect phase" in sum-product alg.

Let $\alpha_t(z_t) \triangleq p(\bar{x}_t | z_t)$

$$\alpha_t = \alpha_t \odot (A \alpha_{t-1})$$

Hadamard product

initialization

$$\alpha_1(z_1) = p(z_1, \bar{x}_1) = p(z_1) p(\bar{x}_1 | z_1)$$

space complexity: $O(k)$ extra storage
 time complexity: $O(tk^2)$

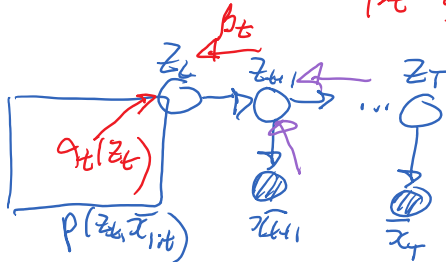
$$\tilde{\alpha}_t(z_t) \triangleq p(z_t | \bar{x}_{1:t}) \text{ "jointing distribution"}$$

$$= \frac{\alpha_t(z_t)}{\sum_{z'} \alpha_t(z')} = \sum_{z'} p(z_t, \bar{x}_{1:t}) = p(\bar{x}_{1:t})$$

"evidence prob."

17h38

β -recursion (smoothing)



$$p(z_t, \bar{x}_{1:t}) = \int_{z_{t-1}} \alpha_t(z_t) M_{z_{t+1} \rightarrow z_t}(z_t) \triangleq \beta_t(z_t)$$

$$M_{z_{t+1} \rightarrow z_t}(z_t) = \sum_{z_{t+1}} p(z_{t+1}|z_t) p(\bar{x}_{t+1}|z_{t+1}) M_{z_{t+2} \rightarrow z_{t+1}}(z_{t+1})$$

$$\beta_t(z_t) = \sum_{z_{t+1}} p(z_{t+1}|z_t) p(\bar{x}_{t+1}|z_{t+1}) \beta_{t+1}(z_{t+1})$$

β -recursion (aka. backward recursion)

turns out that $\beta_t(z_t) \triangleq p(\bar{x}_{t+1:T} | z_t)$

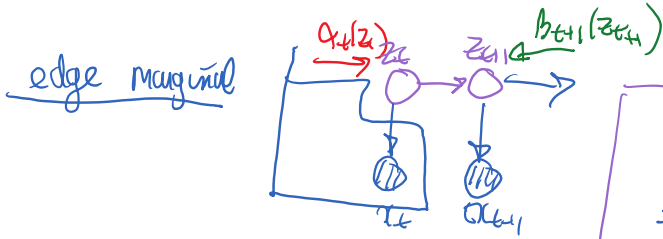
why?

$$p(z_t, \bar{x}_{1:T}) = p(\bar{x}_{1:T} | z_t) p(z_t) \stackrel{\text{CI}}{=} p(\bar{x}_{1:T} | z_t)$$

initialization: $\beta_T(z_T) = 1 \quad \forall z_T$

$$p(z_t, \bar{x}_{1:T}) = p(\bar{x}_{1:T} | z_t) p(z_t)$$

$$\stackrel{\text{def}}{=} p(\bar{x}_{t+1:T} | z_t) \cdot \underbrace{p(\bar{x}_{1:t} | z_t) p(z_t)}_{\alpha_t(z_t)}$$



$$p(z_t, z_{t+1}, \bar{x}_{1:T}) = \alpha_t(z_t) \beta_{t+1}(z_{t+1}) p(z_{t+1} | z_t) p(\bar{x}_{t+1} | z_{t+1})$$

numerical stability trick:

issue: α_t & β_t can easily go to $1e-100$

two possibilities:

a) (general) store $\log(\alpha_t)$ instead ($a_i > 0$)

$$\log\left(\sum_i a_i\right) = \log\left(\tilde{a} \sum_i \frac{a_i}{\tilde{a}}\right)$$

call $\tilde{a} \equiv \max_i a_i$
 $\tilde{a}_{\max} \equiv \max_i \max_i a_i$

$$= \log(\tilde{a}) + \log\left(1 + \sum_{j \neq i_{\max}} \exp(\log(a_j) - \log(\tilde{a}))\right)$$

b) normalize the message

• α -recursion, use $\tilde{\alpha}_t(z_t) = p(z_t | \bar{x}_{1:t})$

before, $\alpha_t = \alpha_t \odot A \alpha_{t-1}$

$$\tilde{\alpha}_t = \frac{\alpha_t \odot A \tilde{\alpha}_{t-1}}{\sum_z (\quad)} \triangleq C_t$$

you can show that

$$C_t = \sum_{z_t} (\alpha_t \odot A \tilde{\alpha}_{t-1}) / z_t = p(\bar{x}_t | \bar{x}_{1:t-1})$$

$$p(\bar{x}_{1:T}) = \prod_{t=1}^T p(\bar{x}_t | \bar{x}_{1:t-1}) = \prod_{t=1}^T C_t$$

• β -recursion:

define $\tilde{\beta}_t(z_t) = \frac{p(\bar{x}_{t+1:T} | z_t)}{p(\bar{x}_{t+1:T} | \bar{x}_{1:t})}$

note: $\sum_{z_t} \tilde{\beta}_t(z_t) \neq 1$
 not norm

exercise: derive $\tilde{\beta}$ -recursion

ML for HMM

• some parametric model for dist. on x_t
 $n(x_t | z_t = k) = \xi(x_t | M_{ik}) \quad 0,0$ (transition on x_t)

- suppose $p(x_t | z_t = k) = f(x_t | \mu_k)$ e.g. Gaussian on x_t
- $p(z_{t+1} = i | z_t = j) = A_{ij}$ $M = (M_k)_{k=1}^K$
- $p(z_1 = i) = \pi_i$

$$\Theta = (\mu, A, \pi)$$

want to estimate $\hat{\mu}, \hat{A}, \hat{\pi}$ by MLE from data $\mathcal{X} = (x^{(i)})_{i=1}^N$
 $x^{(i)} = x_{1:T_t}^{(i)}$

→ use EM: s^{th} iteration

E step: $q_{s+1}(z) = p(z | \mathcal{X}, \Theta^{[s]})$

M step: $\Theta^{[s+1]} = \underset{\Theta \in \Theta}{\text{argmax}} \mathbb{E}_{q_{s+1}} [\log p(x, z | \Theta)]$

complete log-likelihood

$$\log p(x, z | \Theta) = \sum_{i=1}^N \left(\underbrace{\log p(z_1^{(i)})}_{\sum_k z_{1,k}^{(i)} \log \pi_k} + \sum_{t=1}^{T_t} \underbrace{\log p(\bar{x}_t^{(i)} | z_t^{(i)})}_{\sum_k z_{t,k}^{(i)} \log f(\bar{x}_t^{(i)} | \mu_k)} + \sum_{t=2}^{T_t} \underbrace{\log p(z_t^{(i)} | z_{t-1}^{(i)})}_{\sum_{l,m} z_{t-1,l}^{(i)} z_{t,m}^{(i)} \log A_{l,m}} \right)$$

$\mathbb{E}_{q_{s+1}} [\log p(x, z | \Theta)] = \dots$

$\mathbb{E}_{q_{s+1}} [z_{t,k}^{(i)}] = q_{s+1}(z_{t,k}^{(i)} = 1) \triangleq \gamma_{t,k}^{(i)}$

smoothing dist: $p(z_{t,k}^{(i)} = 1 | \bar{x}_{1:T_t}^{(i)}; \Theta^{[s]})$
 (use α - β recursion)

$q_{s+1}(z_{t,e}^{(i)} = 1, z_{t-1,m}^{(i)} = 1)$
 $= p(z_{t,e}^{(i)} = 1, z_{t-1,m}^{(i)} = 1 | \bar{x}_{1:T_t}^{(i)}, \Theta^{[s]}) \triangleq \gamma_{t,e,m}^{(i)}$
 ↑ smoothing edge marginal $\ell \left(\begin{matrix} m \\ \vdots \\ e \end{matrix} \right)$

maximize with respect to Θ :

$$\hat{\pi}_k^{[s+1]} = \frac{\sum_{i=1}^N \gamma_{1,k}^{(i)}}{\sum_{i=1}^N \sum_{k=1}^K \gamma_{1,k}^{(i)}} \quad \left. \vphantom{\sum_{i=1}^N} \right\} N$$

$$\hat{A}_{l,m}^{[s+1]} = \frac{\sum_{i=1}^N \sum_{t=2}^{T_t} \gamma_{t,l,m}^{(i)}}{\sum_{u=1}^K \left(\sum_{i=1}^N \sum_{t=2}^{T_t} \gamma_{t,u,m}^{(i)} \right)}$$

$\hat{\mu}_k \rightarrow \text{soft-count MLE}$

e.g. Gaussians similar to GMM ("weighted empirical mean" with weights $\gamma_{t,k}^{(i)}$)

$$\hat{\mu}_k = \frac{\sum_{i=1}^N \sum_{t=1}^{T_t} \gamma_{t,k}^{(i)} x_t^{(i)}}{\sum_{i=1}^N \sum_{t=1}^{T_t} \gamma_{t,k}^{(i)}}$$

• variable to compute $\mu(\bar{x}_{1:T_t})$

• Viterbi
(max product) \rightarrow compute $\underset{Z_{1:T}}{\text{argmax}} p(Z_{1:T} | \bar{x}_{1:T})$