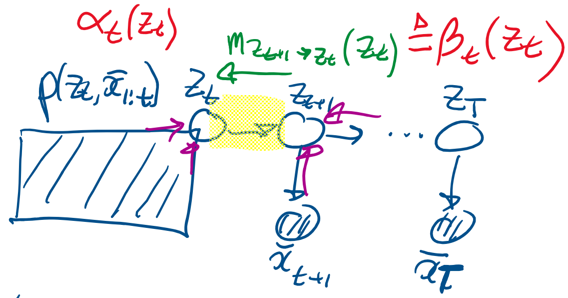
Lecture 16 - EM for HMM; info theory

Thursday, November 2, 2023   2:38 PM

today: · EM for HMM
· info theory & KL

β-recursion: (smoothing)



$$p(z_t, \bar{x}_{1:T}) = \frac{1}{Z} \alpha_t(z_t) \underbrace{m_{z_{t+1}\to z_t}(z_t)}_{\triangleq \beta_t(z_t)}$$

$$\underbrace{m_{z_{t+1}\to z_t}(z_t)}_{\beta_t(z_t)} = \sum_{z_{t+1}} p(z_{t+1}|z_t)\, p(\bar{x}_{t+1}|z_{t+1})\, \overbrace{m_{z_{t+2}\to z_{t+1}}(z_{t+1})}^{\beta_{t+1}(z_{t+1})}$$

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

β-recursion (aka. backward recursion)

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

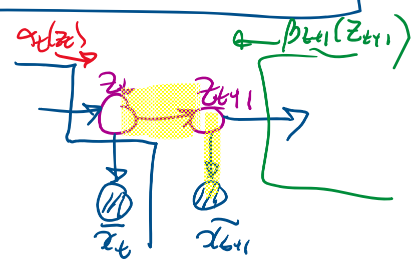
why? $p(z_t, \bar{x}_{1:T}) = p(\bar{x}_{1:T}|z_t)\, p(z_t)$

$$\overset{C.I.}{=} p(\bar{x}_{t+1:T}|z_t) \underbrace{p(\bar{x}_{1:t}|z_t)\, p(z_t)}_{\alpha_t(z_t)}$$

$\Rightarrow \beta_t(z_t)$

$\boxed{\text{initialization } \beta_T(z_T) = 1 \ \forall z_T}$

edge marginal:



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

numerical stability trick:

numerical stability trick:

issue: α_t & β_t can easily go to 1e-1000

two possibilities:

a) (general) store $\log(\alpha_t)$ instead

$$\log\left(\sum_i a_i\right) = \log\left(\tilde{a}\left(\sum_i \frac{a_i}{\tilde{a}}\right)\right) \qquad (a_i > 0)$$

call $\tilde{a} \triangleq \max_i a_i$ $\qquad = \log \tilde{a} + \log\left(\sum_i \frac{a_i}{\tilde{a}}\right)$

$i_{max} \triangleq \arg\max_i a_i$ $\qquad \log\left(\sum_i \exp(\log a_i - \log(\tilde{a}))\right)$

$\log(\tilde{a}) = \log(a_{i_{max}}) = \max_i \log(a_i)$

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

b) <u>normalize the message</u>

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

before $\alpha_t = O_t \odot A\alpha_{t-1}$ $\qquad \tilde{\alpha}_t = \dfrac{O_t \odot A\tilde{\alpha}_{t-1}}{\sum_{z_t}(\quad " \quad)} \Big\} \triangleq c_t$

you can show that $c_t = \sum_{z_t} (O_t \odot A\tilde{\alpha}_{t-1})(z_t)$

$= p(\bar{x}_t \mid \bar{x}_{1:t-1})$

$$p(\bar{x}_{1:T}) = \prod_{t=1}^{T} p(\bar{x}_t \mid x_{1:t-1}) = \prod_{t=1}^{T} c_t$$

- β-recursion:

define $\tilde{\beta}_t(z_t) = \dfrac{\overbrace{\beta_t(z_t)}^{p(\bar{x}_{t+1:T} \mid z_t)}}{p(\bar{x}_{t+1:T} \mid \bar{x}_{1:t})} \Big\} \prod_{u=t+1}^{T} c_u$ $\qquad$ note: $\sum_{z_t} \tilde{\beta}_t(z_t) \neq 1$

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

15h21

<u>ML for HMM</u>

for some parametric model for distr. on x_t

ML for HMM

for some parametric model for distr. on x_t

- suppose $p(x_t|z_t=k) = f(x_t|\eta_k)$ e.g. Gaussian on x_t

$\eta = (\eta_k)_{k=1}^{K}$

- $p(z_{t+1}=i \mid z_t=j) = A_{ij}$
- $p(z_1=i) = \pi_i$

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

want to estimate $\hat{\eta}, \hat{A}$ & $\hat{\pi}$ by ML from data $x = (x^{(i)})_{i=1}^{N}$

$x^{(i)} = x^{(i)}_{1:T_i}$

$\rightsquigarrow$ use EM at s^{th} iteration

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

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

complete log-likelihood

$$\log p(x,z|\Theta) = \sum_{i=1}^{N}\Bigg[\underbrace{\left(\log p(z_1^{(i)})\right)}_{\sum_k z^{(i)}_{1,k}\log \pi_k} + \sum_{t=1}^{T_i}\underbrace{\log p(x_t^{(i)}|z_t^{(i)})}_{\sum_k z^{(i)}_{t,k}\log f(x_t^{(i)}|\eta_k)} + \sum_{t=2}^{T_i}\underbrace{\log p(z_t^{(i)}|z_{t-1}^{(i)})}_{\sum_{l,m} z^{(i)}_{t,l}\, z^{(i)}_{t-1,m}\log A_{lm}}\Bigg]$$

huge variable

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

$\mathbb{E}_{q_{s+1}}[z^{(i)}_{t,k}] = q_{s+1}(z^{(i)}_{t,k}=1) \triangleq \gamma^{(i)}_{t,k}$ ← obtained using α-β recursion on i^{th} time series

smoothing dist: $p(z^{(i)}_{t,k}=1 \mid x^{(i)}_{1:T_i}; \Theta^{[s]})$

$q_{s+1}(z^{(i)}_{t,l}=1, z_{t-1,m}=1) = p(z^{(i)}_{t,l}=1, z^{(i)}_{t-1,m}=1 \mid x^{(i)}_{1:T_i}; \Theta^{[s]})$

$\triangleq \gamma^{(i)}_{t,l,m}$

smoothing edge marginal prob. in HMM (α-β recursion

A matrix

$$\begin{array}{cc} & m \\ l & \left(\ \boxed{\ }\ \right) \end{array}$$

note: you should have

$\sum_l \gamma^{(i)}_{t,l,m} = \gamma^{(i)}_{t-1,m}$

max. with respect to Θ:

$\hat{\pi}_k^{[s+1]} = \sum_{i=1}^{N}\gamma^{(i)}_{1,k}$

$\hat{A}^{[s+1]}_{l,m} = \sum_{i=1}^{N}\sum_{t=2}^{T_i}\gamma^{(i)}_{t,l,m}$

$\hat{\eta}_k \rightarrow$ soft count ML

$$\hat{\pi}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{1,k}^{(i)}}{\underbrace{\sum_{i=1}^{N} \underbrace{\sum_{k=1}^{K} \gamma_{1,k}^{(i)}}_{1}}_{N}}$$

$$\hat{A}_{l,m} = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T_i} \gamma_{t,l,m}^{(i)}}{\sum_{u} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \gamma_{t,u,m}^{(i)} \right)}$$

$n_k \to$ soft count ML

$\to (\hat{\mu}_k, \hat{\Sigma}_k)$ for Gaussian

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

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

Viterbi to compute $\underset{z_{1:T_i}}{\operatorname{argmax}}\ p(z_{1:T_i} \mid x_{1:T_i})$

(max product)

Information theory

KL divergence: for discrete dist. p & q

$$\boxed{KL(p \| q) \triangleq \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right] = \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)}}$$

$0 \cdot \log 0 = 0$

$\left(\lim_{x \to 0^+} x \log x = 0\right)$

[if $\exists x$ s.t. $q(x) = 0$ but $p(x) \neq 0$

$\underbrace{-p(x)}_{\text{not zero}} \underbrace{\log q(x)}_{-\infty} = +\infty$]

$KL = \infty$ if $\sup(p) \not\subseteq \sup(q)$

motivation from density estimation

recall statistical decision theory

(statistical) loss $L(p_{\text{world}}, a)$ — here, estimation of dist., say $\hat{q}$

standard (MLE) loss is log-loss $L(p_\epsilon, \hat{q}) = \underbrace{\mathbb{E}_{X \sim p_\epsilon}\left[-\log \hat{q}(X)\right]}_{\text{cross-entropy}}$

if use $\hat{q} = p_\epsilon$, then get $\sum_{x \in \Omega_X} -p_\epsilon(x) \log p_\epsilon(x) \triangleq \underbrace{H(p_\epsilon)}_{\text{entropy of } p_\epsilon}$

If use $q = p_\theta$ then get $\boxed{\sum_{x\in\Omega_X} -p_\theta(x)\log p_\theta(x) = H(p_\theta)}$ entropy of p_θ

excess loss for action $a = \hat{q}$

$$L(p,\hat{q}) - \underbrace{\min_q L(p,q)}_{L(p,p)} = \sum_{x\in\Omega} -p_\theta(x)\log\frac{\hat{q}(x)}{p_\theta(x)} = KL(p\|\hat{q})$$

<u>coding theory</u>:

use length of code $\propto -\lg p(x)$

$\lg \triangleq \log_2 \rightarrow$ "bits"

$\log_e \rightarrow$ "nats"

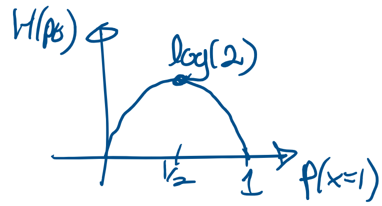
expected length of code : $\sum_x p(x)(-\lg p(x))$ (entropy measured in bits)

KL divergence $\rightarrow$ interpreted as excess length cost (in terms of length of code) to use dist q to design code vs optimal dist (true p)

<u>example</u>:

<u>entropy of a Bernoulli</u>

$-p\log p - (1-p)\log(1-p)$



<u>entropy of a uniform dist. on k states</u>

$$\sum_{x=1}^{k} -\frac{1}{k}\log\left(\frac{1}{k}\right) = \log k$$

(max entropy dist. over k states)

<u>properties of KL</u>

- $KL(p\|q) \geq 0$ $\leftarrow$ to show this, use Jensen's ineq. $f(\mathbb{E}X) \leq \mathbb{E}f(X)$ when f is convex

- KL is <u>strictly</u> convex in each of its arguments

i.e. $KL(p\|\cdot) : \Delta_k \subseteq \mathbb{R}^k \rightarrow \mathbb{R}$ strictly convex

$KL(\cdot\|q)$

- <u>not symmetric</u> $KL(p\|q) \neq KL(q\|p)$ in general

$KL(p\|p) = 0 \quad \forall p \in \Delta_k$

symmetrized $\frac{1}{2}(KL(p\|q) + KL(q\|p))$

in general

symmetrized version $\frac{1}{2}\left(KL(p\|q) + KL(q\|p)\right)$