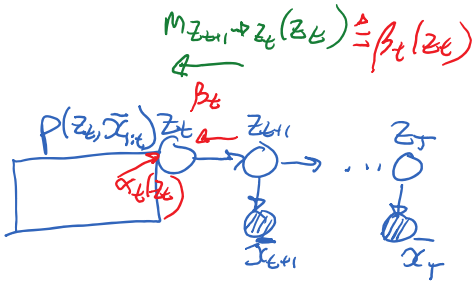
Lecture 16 - EM for HMM; info theory

Friday, November 4, 2022 2:56 PM

today: • EM for HMM
• info. theory & KL

β-recursion (smoothing)



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

$Z = 1$

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

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

β-recursion (aka. backward recursion)

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

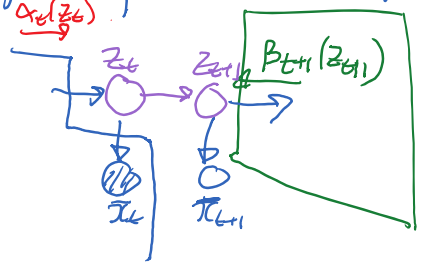
why? $p(z_t, \bar{x}_{1:T}) = p(\bar{x}|z_t)\,p(z_t)$

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

$\Rightarrow \beta_t(z_t)$

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

edge marginals



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

numerical stability trick:

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

two possibilities:

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

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

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

call $\tilde{a} \triangleq \max_i a_i$ $\quad i_{max} \triangleq \arg\max_i a_i$

$\log(\tilde{a}) \triangleq \max_i \log(a_i)$

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

b) normalize the message

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

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

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

you can show that $c_t = \sum_{z_t} (O_t \odot A\tilde{\alpha}_{t-1})(z_t) = p(\bar{x}_t \mid \bar{x}_{1:t-1})$

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

- 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

15h50

ML for HMM

- suppose $p(x_t \mid z_t = k) = f(x_t \mid \eta_k)$ — same parametric model for dist. on x_t (e.g. Gaussian on x_t) $\quad \eta = (\eta_k)_{k=1}^K$
- $p(z_{t+1} = i \mid z_t = j) = A_{ij}$
- $p(z_1 = i) = \pi_i$ $\qquad \Theta = (\eta, A, \pi)$

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 $\qquad$ at s^{th} iteration

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

M step: $\quad \hat{\Theta}^{[s+1]} = \arg\max_{\Theta \in \mathbb{H}} \mathbb{E}_{q_{s+1}}[\log p(x, z \mid \Theta)]$

complete log-likelihood

$$\log p(x, z \mid \Theta) = \sum_{i=1}^N \left(\log p(z_1^{(i)}) + \sum_t^{T_i} \log p(x_t^{(i)} \mid z_t^{(i)}) + \sum_t^{T_i} \log p(z_t^{(i)} \mid z_{t-1}^{(i)})\right)$$

Complete log-likelihood

$$\log p(x, z | \Theta) = \sum_{i=1}^{N} \left(\log p(z_1^{(i)}) + \sum_{t=1}^{T_i} \log p(x_t^{(i)} | z_t^{(i)}) + \sum_{t=2}^{T_i} \log p(z_t^{(i)} | z_{t-1}^{(i)}) \right)$$

(huge variable)

$$\sum_k z_{1,k}^{(i)} \log \pi_k \qquad \sum_k z_{t,k}^{(i)} \log f(x_t^{(i)} | \mu_k) \qquad \sum_{l,m} z_{t,l}^{(i)} z_{t-1,m}^{(i)} \cdot \log A_{lm}$$

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

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

smoothing dist $p(z_{t,k=1}^{(i)} | x_{1:T_i}^{(i)}; \Theta^{[s]})$ — obtained via α-β recursion

$$q_{s+1}(z_{t,l}^{(i)} = 1, z_{t-1,m}^{(i)} = 1) = p(z_{t,l}^{(i)} = 1, z_{t-1,m}^{(i)} = 1 | x_{1:T_i}^{(i)}; \Theta^{[s]}) \triangleq \gamma_{t,l,m}^{(i)}$$

A matrix (rows l, columns m)

note: you should have: $\sum_l \gamma_{t,l,m}^{(i)} = \gamma_{t,m}^{(i)}$

smoothing edge marginal in HMM (α-β recursion)

maximize with respect to Θ:

$$\hat{\pi}_k^{[s+1]} = \frac{\sum_{i=1}^{N} \gamma_{1,k}^{(i)}}{\sum_{i=1}^{N} \sum_{l} \gamma_{1,l}^{(i)}} \quad \left(\text{denominator} = N, \ \sum_l \gamma_{1,l}^{(i)} = 1\right)$$

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

$\hat{n}_k \rightarrow$ soft counts ML

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

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

Viterbi to compute $\arg\max_{z_{1:T_i}} p(z_{1:T_i} | x_{1:T_i}^{(i)})$
(max product)

Information theory

KL divergence: for discrete dist. p & q

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

$0 \cdot \log 0 = 0$

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

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

$-p(x) \log q(x) = +\infty$] ($p(x)$ not zero, $\log q(x) = -\infty$)

if support of $p \not\subseteq$ support of q

$\Rightarrow KL(p\|q) = +\infty$

motivation from density estimation

recall statistical decision theory

(statistical) loss $L(p_{\theta}, \hat{q})$ — world: p_{θ}; here, estimation of dist., say $\hat{q}$

Standard (MLE) loss is log-loss $L(p_{\theta}, \hat{q}) = \mathbb{E}_{x \sim p_{\theta}}[-\log \hat{q}(x)]$

if use $\hat{q} = p_{\theta}$, then get $\sum_{x \in \Omega_X} -p_{\theta}(x) \log p_{\theta}(x) \triangleq H(p_{\theta})$ (cross-entropy)

entropy of p_{θ}

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

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

$\min_q L(p,q) = L(p,p)$

$\log \triangleq \log_2 \rightarrow$ "bits"

$\log_e \rightarrow$ "nats"

coding theory:

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

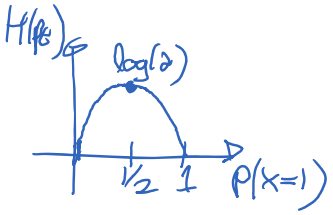
expected length of code: $\sum_x p(x)(-\log p(x))$ (entropy measured in bits)

KL divergence → interpreted as excess length of code (in terms of length of code) to use dist. q to design code vs. optimal dist (true p)

example:

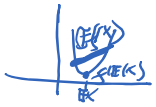
entropy of a Bernoulli:

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



entropy for a uniform dist on k states: $\sum_{x=1}^{k} -\frac{1}{k}\log\left(\frac{1}{k}\right) = \log k$

(max. entropy dist. over k states)



properties of KL

- $KL(p\|q) \geq 0$ ← to show this, use Jensen's inequality $f(\mathbb{E}X) \leq \mathbb{E}f(X)$ when f is convex
- KL is strictly convex in each of its arguments i.e. $KL(p\|\cdot): \mathbb{R}^K \rightarrow \mathbb{R}$

$KL(\cdot \| q)$ Strictly convex fct.

- not symmetric $KL(p\|q) \neq KL(q\|p)$ in general

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

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