Lecture 18 - EM in HMM; info theory

Friday, November 6, 2020 13:24

today : · EM for HMM
· info. theory & KL max / entropy

Continue α-recursion:

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

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

let $o_t(z_t) \triangleq p(\bar{x}_t \mid z_t)$

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

$\odot$ ← Hadamard product

initialization:

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

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

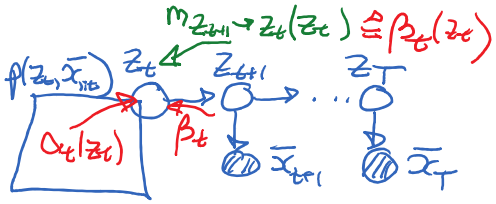
$$= \frac{\alpha_t(z_t)}{\sum_{z_t} \alpha_t(z_t)} \Big\} = \sum_{z_t} p(z_t, \bar{x}_{1:t}) = p(\bar{x}_{1:t})$$

"evidence probability"

space complexity: $O(k)$ extra storage

time complexity: $O(tk^2)$

β-recursion (smoothing)



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

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

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

β-recursion (aka backward recursion)

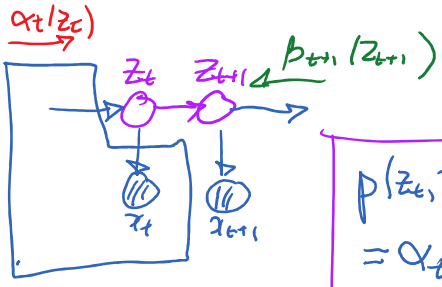
turns out that $\boxed{\beta_t(z_t) \triangleq p(\bar{x}_{t+1:T} | z_t)}$ why?

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

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

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

edge marginal



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

Numerical stability trick:

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

two possibilities:

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

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

call $\tilde{a} \triangleq \max_i a_i$, $i_{max} \triangleq \operatorname{argmax}_i a_i$

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

b) normalize the message

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

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

$$\boxed{\tilde{\alpha}_t = \frac{o_t \odot A \tilde{\alpha}_{t-1}}{\underbrace{\sum_{z_t} (\quad)}_{\triangleq c_t}}}$$

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

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

- β-recursion:

define $\tilde{\beta}_t(z_t) = \dfrac{\beta_t(z_t)}{p(\bar{x}_{t+1:T} | x_{1:t})} \prod_{u=t+1}^{T} c_u$ where $p(\bar{x}_{t+1:T} | z_t)$

note: $\sum_{z_t} \tilde{\beta}_t(z_t) \neq 1$

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

14h24

ML for HMM

- suppose $p(x_t | z_t = k) = f(x_t | \eta_k)$ — some parametric model for dist. on x_t, e.g. Gaussian on x_t
- $p(z_{t+1} = i \mid z_t = j) = A_{ij}$
- $p(z_1 = i) = \pi_i$

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

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

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

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

$\rightsquigarrow$ use EM

E step: (s^{th} iteration)

$$q_{s+1}(z) = p(z | x, \theta^{[s]})$$

M step:

$$\theta^{[s+1]} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; \mathbb{E}_{q_{s+1}}[\log p(x, z)]$$

Complete log-likelihood:

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

$\log p(z_1^{(i)}) = \sum_k z_{1,k}^{(i)} \log \pi_k$

$\log p(x_t^{(i)} | z_t^{(i)}) = \sum_k z_{t,k}^{(i)} \log f(x_t^{(i)} | \eta_k)$

$\log p(z_t^{(i)} | z_{t-1}^{(i)}) = \sum_{l,m} z_{t,l}^{(i)} z_{t-1,m}^{(i)} \log A_{lm}$

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

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

smoothing dist $p(z_{t,k}^{(i)} = 1 \mid x_{1:T_i}^{(i)}; \theta^{[s]})$

α-β recursion

$q_{s+1}(z_{t,l}^{(i)} = 1, z_{t-1,m}^{(i)} = 1)$

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

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

smoothing edge marginal $\left(l \to \boxed{\hat{\ }}\right)$

maximize with respect to θ:

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

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

$\hat{n}_k \to$ soft-counts n_L

e.g. Gaussians

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

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

correction 2020/11/19 from original scribbles

Viterbi (max product) to compute $\operatorname{argmax}_{z_{1:T}} p(z_{1:T} \mid x_{1:T})$

Information theory

KL divergence : for discrete dist. p & q

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

$0 \cdot \log 0 = 0$

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

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

$-p(x) \underbrace{\log q(x)}_{-\infty} = +\infty$]

if support of $p \not\subseteq$ support of $q \Rightarrow KL(p \| q) = +\infty$

motivation from density estimation

recall statistical decision theory world

here, estimation of dist., say $\hat{q}$

(statistical) loss $L(p_\theta, q)$

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 $\sum_{x \in \mathcal{X}} -p_\theta(x) \log p_\theta(x) \triangleq H(p_\theta)$ (cross-entropy)

entropy of p_θ

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

$$L(p, \hat{q}) - \min_q L(p,q) = \sum_x p(x) \log \frac{\hat{q}(x)}{p(x)}$$

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

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

Coding theory:

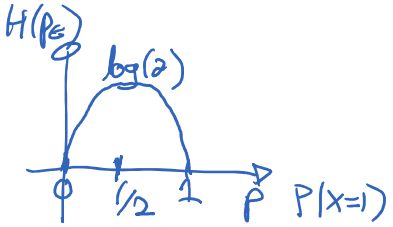
use length of code $\propto -\lg p(x)$

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

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

Example:

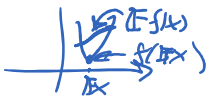
entropy for a Bernoulli: $-p \log p - (1-p)\log(1-p)$



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

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

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