Lecture 15 - scribbles

Tuesday, October 31, 2017
14:30

today: finish HMM + EM
information theory & KL

HMM continued:

recall: $\alpha_t(z_t) \triangleq p(z_t, \bar{x}_{1:t})$

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

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

O_t: observation prob vector for z_t $\quad p(\bar{x}_t|\cdot)$

numerical trick & issue: α_t can easily be $1e{-}100$

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

$$a+b = \exp(\log(a)) + \exp(\log(b))$$

say $a \geq b$

$$= a(1 + \exp(\log(b) - \log(a)))$$

$$\log(a+b) = \log(a) + \log(1 + \exp(\log(b) - \log(a)))$$

in general $\sum_i a_i$

first pick $\max_i a_i$

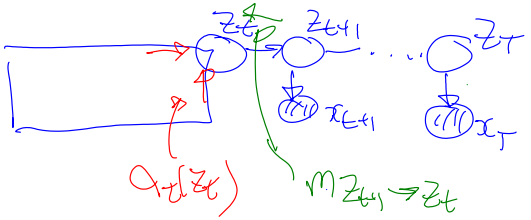
b) normalize the messages

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

before $\alpha_t = O_t \odot A\alpha_{t-1}$ $\quad \tilde{\alpha}_t = O_t \odot A\tilde{\alpha}_{t-1}$ you can show that

before $\alpha_t = O_t \odot A\alpha_{t-1}$ $\quad \tilde{\alpha}_t = \dfrac{O_t \odot A\tilde{\alpha}_{t-1}}{\sum_{z_t}[\;\dots\;]}$ you can show that normalization constant is $p(\bar{x}_t \mid x_{1:t-1})$

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



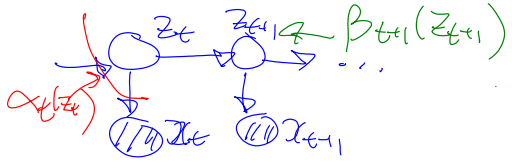
$$m_{z_{t+1}\to z_t}(z_t) = \sum_{z_{t+1}} p(z_{t+1}\mid z_t)\, p(\bar{x}_{t+1}\mid z_{t+1})\, m_{z_{t+2}\to z_{t+1}}(z_{t+1})$$

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

β-recursion (backward recursion)

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

edge marginal:



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

ML for HMM:

- suppose $p(x_t \mid z_t = k) = f(x_t \mid \eta_k)$ $\quad \eta = (\eta_k)_{k=1}^K$
- $p(z_{t+1} = i \mid z_t = j) = A_{ij}$
- $p(z_1 = i) = \pi_i$ $\quad \Theta = (\eta, A, \pi)$

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

$\rightarrow$ use EM $\quad$ sth iteration

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

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

complete log-likelihood:

$$\log p(x, z \mid \theta) = \sum_{i=1}^{N} \left[\log p(z_1^{(i)}) + \sum_{t=1}^{T} \log p(x_t^{(i)} \mid z_t^{(i)}) + \sum_{t=2}^{T} \log p(z_t^{(i)} \mid z_{t-1}^{(i)}) \right]$$

x: huge variables

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

$\sum_k z_{t,k}^{(i)} \log f(x_t^{(i)} \mid m_k)$

$\sum_{\ell,m} z_{t,\ell}^{(i)} z_{t-1,m}^{(i)} \log A_{\ell m}$

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

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

$q_{s+1}(z_{t,\ell}^{(i)} = 1, z_{t-1,m}^{(i)} = 1) = p(z_{t,\ell}^{(i)} = 1, z_{t-1,m} = 1 \mid x_{1:T}^{(i)}, \theta^{[s]})$

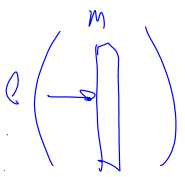
smoothing marginal

$\gamma_{t,\ell,m}^{(i)}$

maximizing with respect to θ:

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

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



"Baum-Welch" alg. — forward-backward alg. (a-p recursion / sum-product) + EM for HMM

https://en.wikipedia.org/wiki/Baum%E2%80%93Welch_algorithm

note:

$$\boxed{\beta_t(z_t) = p(x_{t+1:T} \mid z_t)}$$

$$p(z_t, x_{1:T}) = \alpha_t(z_t)\beta_t(z_t)$$

$$\Rightarrow \beta_t(z_t) = \frac{p(z_t, x_{1:T})}{p(z_t, x_{1:t})\alpha_t(z_t)} = \frac{p(z_t, x_{t+1:T} \mid x_{1:t})\, p(x_{1:t})}{p(z_t \mid x_{1:t})\, p(x_{1:t})}$$

$= p(\bar{x}_{t+1:T} \mid z_t, \bar{x}_{1:t}) \frac{p(z_t \mid x_{1:t})}{p(z_t \mid x_{1:t})}$

$= p(\bar{x}_{t+1:T} \mid z_t)$ by cond. indep. //

Information theory:

KL divergence: for discrete dist. p & q

$$KL(p \| q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$0 \cdot \log 0 = 0$

motivation from density estimation:

recall statistical decision theory

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

action is estimating a distribution say $\hat{q}$

standard (ML) 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 $\sum_{x \in \mathcal{X}} -p(x) \log p(x) \triangleq H(p)$ ("cross-entropy")

entropy of p

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

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

$L(p, p)$

coding theory: use length of code of $-\log p(x)$ (where q drives)

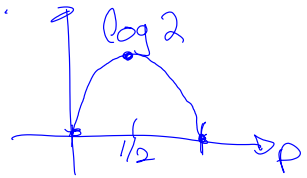
$\log_2 \rightarrow$ "bits" $\quad \log \rightarrow$ "nats"

[minimal expected length]

expected length of code: $\sum_x p(x)\left(-\log q(x)\right)$ (like the entropy)

KL divergence → excess cost (in terms of expected length of code) to use dist. q for coding instead of p

entropy a Bernoulli:



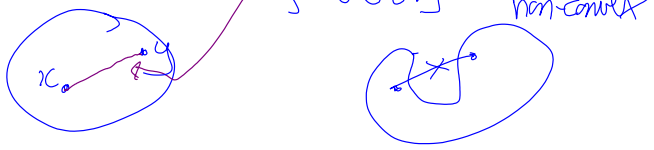
entropy of uniform dist. on k states: $-\sum_x \frac{1}{k}\log\left(\frac{1}{k}\right) = \log k$

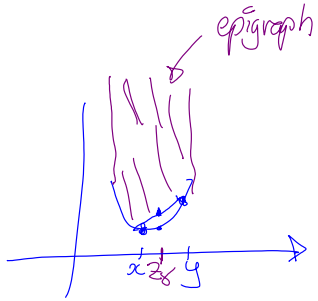
properties of KL:

- $KL(p\|q) \geq 0$
- is strictly convex in each argument i.e. $KL(p\|\cdot)$, $KL(\cdot\|q)$

convexity review:

set A is convex $\Leftrightarrow \forall x,y \in A \quad (1-\gamma)x + \gamma y \in A$ for $\gamma \in [0,1]$





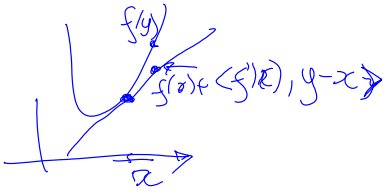
function f is convex $\Leftrightarrow$ it's epigraph is convex

↳ $\{(x,t) : x \in \text{dom}(f), t \geq f(x)\}$

$\Leftarrow$ $f((1-\gamma)x+\gamma y) \overset{\text{convex}}{\leq} (1-\gamma)f(x)+\gamma f(y) \quad \forall \gamma \in [0,1] \quad \forall x,y \in \text{dom} f$

(z_γ; $<$ "strictly convex")

if f is differentiable, $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle \quad \forall y,x$



recall Jensen: $f(\mathbb{E}X) \leq \mathbb{E}f(X)$ (if convex)

ML and KL minimization:

$\{p_\theta\}_{\theta \in \Theta}$ parametric family

empirical dist.

then ML for $\theta \iff \min_{\theta \in \Theta} KL(\hat{p}_n \| p_\theta)$

proof: $KL(\hat{p}_n \| p_\theta) = \sum_x \hat{p}_n(x) \log \frac{\hat{p}_n(x)}{p_\theta(x)}$

$= -H(\hat{p}_n) - \sum_x \hat{p}_n(x) \log p_\theta(x)$

$\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^n \delta(x, x^{(i)})$ (Kronecker-delta)

$= -H(\hat{p}_n) - \frac{1}{n} \sum_{i=1}^n \log p_\theta(x^{(i)})$

$-H(\hat{p}_n)$: constant

$\log \prod_{i=1}^n p_\theta(x^{(i)})$

//

next: max entropy

next: max entropy