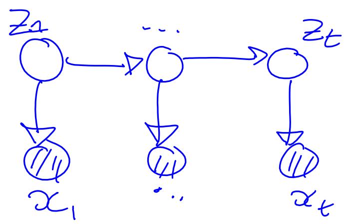
Lecture 14 - scribbles

Tuesday, October 18, 2016
14:50

today: · HMM
· max entropy, exp family
and beautiful duality

HMM: hidden Markov model



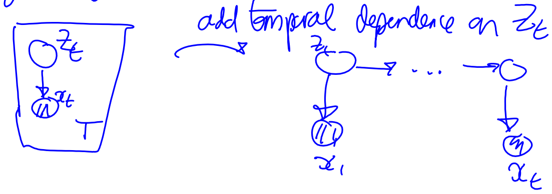
$z_t \in \{1, \dots, K\}$ discrete

x_t < cts. → e.g. speech signal
discrete → DNA sequence

speech recognition x_t → speech signal
z_t → phonemes

HMM → generalization of mixture model

GMM



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 prob. and transition prob are homogeneous i.e. do not depend on t

$$p_t(x_t | z_t) = f(x_t | z_t)$$

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

A (j [ ] } dist. over z)

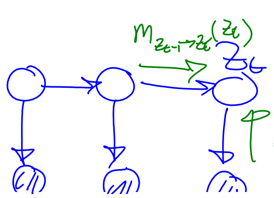
$\sum_i A_{ij} = 1$

inference tasks:

prediction $p(z_t | x_{1:t-1})$ "where next?"

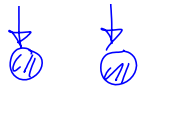
filtering $p(z_t | x_{1:t})$ "where now?"

smoothing $p(z_t | x_{1:T})$ "where in the past"
$T > t$



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

$m_{x_t \to z_t}(z_t) = \frac{1}{Z} \, m_{x_t \to z_t}(z_t) \, m_{z_{t-1} \to z_t}(z_t)$



$$m_{x_t \to z_t}(z_t) = \sum_{x_t} p(x_t \mid z_t)\, \delta(x_t, \bar{x}_t)$$

$$= p(\bar{x}_t \mid z_t)$$

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

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

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

→ Hadamard product

α-recursion, forward recursion like the "collect phase" in sum-product with z_t the root

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

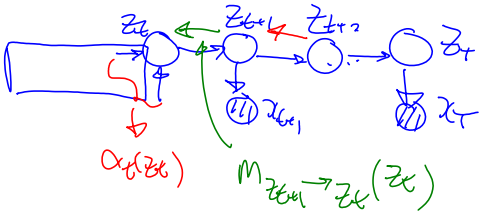
time complexity: $O(t \cdot K^2)$

space complexity $O(K)$ extra ($O(K^2)$ count storing A)

$$\sum_{z_t} \alpha_t(z_t) = p(x_{1:t})$$

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

$Z = 1$ for DGM



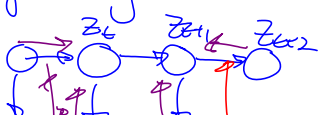
$$m_{z_{t+1} \to z_t}(z_t) = \sum_{z_{t+1}} p(z_{t+1} \mid z_t)\, p(\bar{x}_{t+1} \mid z_{t+1})\, m_{z_{t+2} \to z_{t+1}}(z_{t+1})$$

$$\beta_t(z_t) = \sum \sim \cdot \beta_{t+1}(z_{t+1})$$

initialization: → β-recursion "backward recursion"

$$\beta_T(z_T) = m_{z_{T+1} \to z_T}(z_T) = 1$$

finally, can get marginals on two nodes



$$p(z_t, z_{t+1}, x_{1:T})$$

$= \alpha_t(z_t)\,\beta_{t+1}(z_{t+1})\,p(x_t \mid z_t)\,p(x_{t+1} \mid z_{t+1})\cdot p(z_{t+1} \mid z_t)$

$\alpha_t(z_t)$ $\beta_{t+1}(z_{t+1})$

ML in HMM:

- suppose $p(x \mid z=k) = f(x \mid \theta_k)$
- $p(z_{t+1} = i \mid z_t = j) = A_{ij}$

want to estimate $\hat{A}$ & $\hat{\theta}$ by ML from data

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

EM write complete log-likelihood $\log p(x, z)$ — E step → $q_{s+1}(z) \to p(z \mid x, \theta^{[s]})$ (parameter at s^{th} iteration)

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

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

(huge variables)

$\sum_k z_{1,k}^{(i)} \log \pi_k$ $\quad \sum_k z_{t,k}^{(i)} \log f(x_t^{(i)} \mid \theta_k)$ $\quad \sum_{l,m} z_{t,l}^{(i)} z_{t-1,m}^{(i)} \log A_{l,m}$

$\mathbb{E}_{q_{s+1}}[\log p(x, z)] = \cdots$ $q_{s+1}(z_{1,k}^{(i)} = 1)$ $\to \gamma_{t,k}^{(i)}$ (soft-counts)

$q_{s+1}(z_{t,l}^{(i)} = 1, z_{t-1,m}^{(i)} = 1)$ → smoothing marginals $p(z_t^{(i)} = l, z_{t-1}^{(i)} = m \mid x_{1:T_i}^{(i)}, \theta^{[s]})$

maximizing for θ:

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

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

$l \to m$

KL divergence and density estimation

Statistical decision theory

recall loss: $L(P, a)$

if action is estimate a distribution say $p_{\hat{\theta}}$

standard (ML) loss is log-loss $L(P, \hat{p}_{\hat{\theta}}) = \mathbb{E}_{X \sim P}[-\log p_{\hat{\theta}}(X)]$

cross-entropy

suppose $P = P_{\hat{\theta}}$

Then this loss becomes: $\sum_{x \in \Omega_X} -p(x) \log p(x) \triangleq H(p)$ entropy of p

excess loss for action $a = P_{\hat{\theta}}$,

$$L(P, P_{\hat{\theta}}) - \min_{q \in A} L(P, q) = -\sum_{x \in \Omega_X} p(x) \log \frac{p_{\hat{\theta}}(x)}{p(x)}$$

$$= KL(p \,||\, p_{\hat{\theta}})$$