Lecture 16 - scribbles

Friday, November 8, 2019 13:58

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

HMM continuations:

numerical stability trick:

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

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

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

call $\tilde{a} \triangleq \max_i a_i$

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

$i_{max} \triangleq \operatorname{argmax}_i a_i$

b) normalize the messages:

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

  before, $\alpha_t = o_t \odot A\alpha_{t-1}$ have: $\tilde{\alpha}_t = \underbrace{o_t \odot A\tilde{\alpha}_{t-1}}_{\sum_{z_t}(\;) \triangleq c_t}$

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

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

- β-recursion:

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

  note: $\sum_{z_t} \beta_t(z_t) \neq 1$

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

ML for HMM:

- some parametric model for dist. on x e.g. Gaussian on x_t
- $p(x_t | z_t = k) = C(x_t | \theta_k)$

ML for HMM:

- suppose $p(x_t \mid z_t = k) = f(x_t \mid \eta_k)$ ← some parametric model for dist. on x e.g. Gaussian on x_t

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

- $p(z_{t+1} = i \mid z_t = j) = A_{ij}$
- $p(z_1 = i) = \pi_i$ $\qquad \theta = (\eta, A, \pi)$

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

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

↝ use EM

E step: s^{th} iteration

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

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

complete log-likelihood:

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

- $\log p(z_1^{(i)}) = \sum_k z_{1,k}^{(i)} \log \pi_k$
- $\log p(x_t^{(i)} \mid z_t^{(i)}) = \sum_k z_{t,k}^{(i)} \log f(x_t^{(i)} \mid \eta_k)$
- $\log p(z_t^{(i)} \mid z_{t-1}^{(i)}) = \sum_{l,m} z_{t,l}^{(i)} z_{t-1,m}^{(i)} \log A_{lm}$

(hyper variable)

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

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

smoothing dist. $p(z_t^{(i)} \mid x_{1:T_i}^{(i)}, \theta)$

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

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

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

smoothing edge marginal

$\left(l \to \boxed{}^{m} \right)$

maximize with respect to θ:

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

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

$\hat{\eta}_k$ → soft count ML

e.g. Gaussians, similar to GMM

$\sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_{t,l}^{(i)} \Big] N$ $\quad$ $\sum_{u} \sum_{i=1}^{N} \sum_{t=2}^{T} \gamma_{t,u,m}^{(i)}$

e.g. Gaussians, similar to GMM "weighted empirical mean"

weights $\gamma_{t,k}^{(i)}$

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

Viterbi (max product) to compute $\operatorname{argmax}_{z_{1:T}} p(z_{1:T} | \tilde{x}_{1:T})$

14h55

Information theory

KL divergence: for discrete dist p & q

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

$0 \cdot \log 0 = 0$

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

motivation from density estimation

recall statistical decision theory world has, estimation of dist., say $\hat{q}$

(statistical) loss $L(p_\theta, \hat{a})$

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

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

entropy of p_θ

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

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

$L(p,p)$

Coding theory:

$\lg \triangleq \log_2$ → "bits" $\quad \log_e$ → "nats"

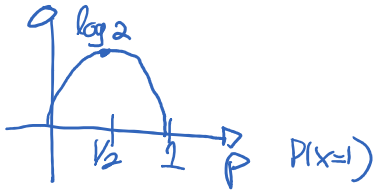
use length of code of $-\log p(x)$

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

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

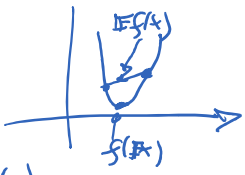
Example

entropy for a Bernoulli:



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

(max. entropy dist. over K-states)



properties of KL:

- $KL(p\|q) \geq 0$ ← to show this, use Jensen's ineq. $f(\mathbb{E}x) \leq \mathbb{E}f(x)$ when f is convex
- KL is strictly convex in each of its arguments i.e. $KL(p\|\cdot): \mathbb{R}^K \to \mathbb{R}$ } strictly convex fct.
- not symmetric $KL(p\|q) \neq KL(q\|p)$ in general $\qquad KL(\cdot\|q)$ $KL(\vec{p}\|\vec{p}) = 0 \;\forall \vec{p} \in \Delta_K$

ML and KL minimization:

$\{p_\theta\}_{\theta\in\Theta}$ parametric family for a discrete observation space

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

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

proof:

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

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

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

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

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

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

Maximum entropy principle:

Idea: consider some subsets of dist. over $\mathcal{X}$
according to some data-driven constraints

get a subset $\mathcal{M} \subseteq \Delta_{|\mathcal{X}|}$

MAXENT principle: pick $\hat{p} \in \mathcal{M}$ which maximizes the entropy

i.e. $\hat{p} = \arg\max_{q \in \mathcal{M}} H(q)$

$= \arg\min_{q \in \mathcal{M}} KL(q \,\|\, \text{uniform})$

$KL(q\|u) = \sum_x q(x) \log \frac{q(x)}{u(x)} = -H(q) + \text{const.}$ ($u(x) = 1/k$ = const.)

"generalized max. entropy" $KL(q\|h_0)$ (h_0 = preferred dist. to bias towards)

* example from Wainwright

$p_L = \frac{3}{4}$ Kangaroos are left-handed

$p_B = \frac{2}{3}$ " drink Sabatt beer

question: how many K. are both L.H. & drink S. beer?

[here: max entropy soln is that $p(B, L) = p_B \cdot p_L$ (indep.)]

* how do we get $\mathcal{M}$?

typically: through empirical "moments"

feature functions $T_1(x), \ldots, T_d(x)$

define $\mathcal{M} = \{q : \mathbb{E}_q[T_j(X)] = \mathbb{E}_{\hat{p}_n}[T_j(X)] \quad j = 1, \ldots, d\}$

define $M = \{ q : \mathbb{E}_q[T_j(X)] = \mathbb{E}_{\hat{p}_n}[T_j(X)] \quad j = 1, \dots, d \}$

$\mathbb{E}_q[T_j(X)]$: model expected feature count

$\mathbb{E}_{\hat{p}_n}[T_j(X)]$: empirical feature count

("moment constraints")

then Max-ENT

$$\min_{q \in \mathbb{R}^{|\mathcal{X}|}} KL(q \,\|\, \text{unif}) \quad \text{s.t. } q \in M, \quad q \in \Delta_{|\mathcal{X}|}$$

$$\sum_x q(x) T_j(x) = \frac{1}{n} \sum_{i=1}^{n} T_j(x^{(i)}) = \alpha_j$$

ie. $\langle \vec{q}, \vec{T}_j \rangle = \alpha_j$

$\rightsquigarrow$ convex opt. problem over $q \in \Delta_{|\mathcal{X}|} \subseteq \mathbb{R}^{|\mathcal{X}|}$

next class: we'll show that Max-ENT is equivalent to max. likelihood in the exp. family with $\vec{T}(x)$ as sufficient statistics