Lecture 10 - scribbles

Friday, October 6, 2017
13:21

today's: finish EM & GMM
directed graphical model

EM continuation:

$$\mathcal{L}(q,\theta) = \mathbb{E}_q\left[\log \frac{p(x,z;\theta)}{q(z)}\right]$$

$$\log p(x;\theta) - \mathcal{L}(q,\theta) = -\mathbb{E}_q\left[\log \frac{p(x,z;\theta)}{q(z)\,p(x;\theta)}\right] \qquad p(z|x;\theta)$$

$$= \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x;\theta)}\right]$$

$$= KL(q(\cdot)\,\|\,p(\cdot|x;\theta)) \quad \text{(KL-divergence)}$$

$\log p(x;\theta) = KL(q\,\|\,p(\cdot|x;\theta)) + \mathcal{L}(q,\theta)$

Operationally:

E step: $q_{t+1} = \underset{q \in \text{all dist. over } z}{\operatorname{argmax}}\ \mathcal{L}(q,\theta_t) = \underset{q}{\operatorname{argmin}}\ KL(q\,\|\,p(\cdot|x;\theta_t))$

$\Rightarrow q_{t+1}(z) = p(z|x;\theta_t)$ [inference]

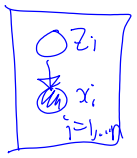
M step $\theta_{t+1} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathcal{L}(q_{t+1},\theta)$

$E_{q_{t+1}}[\log p(x,z;\theta)] + \text{cst.}$

"expected complete log-likelihood"

need to solve another ML problem, but for complete information (i.e. z is known)

for GMM model:



$z_i \overset{iid}{\sim} \text{Mult}(\pi)$

$x_i \mid z_i = j \sim N(\mu_j, \Sigma_j)$

$\theta = (\pi, \{\mu_j\}_{j=1}^K, \{\Sigma_j\}_{j=1}^K)$

notation here $x = x_{1:n}$, $z = z_{1:n}$

complete log-likelihood

$$\log p(x,z;\theta) = \sum_{i=1}^n [\log p(x_i \mid z_i;\theta) + \log p(z_i;\theta)]$$

↓ Gaussian log-likelihood ↓ multinomial

$$\simeq \sum_{i=1}^n \left[\sum_{j=1}^K z_{ij} \log N(x_i \mid \mu_j, \Sigma_j) + \sum_{j=1}^K z_{ij} \log \pi_j\right]$$

$$E_q[\log p(x,z;\theta)] \simeq \sum_{i=1}^n \sum_{j=1}^K E_q[z_{ij}] [\log N(x_i \mid \mu_j, \Sigma_j) + \log \pi_j]$$

$E_q[z_{ij}]$ → $q(z_{ij}=1)$ (marginal distribution)

E step is computing $q_{t+1}(z) \triangleq p(z \mid x; \theta_t)$

$$\propto p(x \mid z;\theta_t)\, p(z;\theta_t)$$

$$\prod_{i=1}^n p(x_i \mid z_i;\theta_t)\, p(z_i;\theta_t)$$

marginal

$$\Rightarrow \quad q_{t+1}(z_i) \propto p(x_i \mid z_i; \Theta_t)\, p(z_i; \Theta_t)$$

$$\qquad\qquad N(x_i \mid \mu_{z_i}, \Sigma_{z_i}) \quad \pi_{z_i}$$

weight $\tau_{i,j}^{t} \triangleq p(z_{ij}=1 \mid x_i, \Theta_t) = q_{t+1}(z_{ij}=1)$

$$= \frac{\pi_j^{(t)} N(x_i \mid \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{l=1}^{k} \pi_l^{(t)} N(x_i \mid \mu_l^{(t)}, \Sigma_l^{(t)})} \quad \begin{matrix} \} \; p(x_i, z_{ij}=1 \mid \Theta_t) \\ \} \; p(x_i \mid \Theta_t) \end{matrix}$$

E step: compute $\tau_{i,j}^{t}$ for $i=1,\dots,n$, $j=1,\dots,k$

M step: $\max_{\{\mu_j, \Sigma_j, \pi_j\}} \sum_{i=1}^{n} \sum_{j=1}^{k} \tau_{ij}^{t} \left[\log p(x_i \mid \mu_j, \Sigma_j) + \log \pi_j\right]$

exercise: $\hat{\pi}_j^{(t+1)} = \frac{\sum_i \tau_{i,j}^{t}}{n}$ ← "soft count"

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{t} x_i}{\left(\sum_{i=1}^{n} \tau_{ij}^{t}\right)}$$

[weighted empirical mean]

(τ_{ij}^{t}: soft cluster assignment)

$$\hat{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{t} (x_i - \hat{\mu}_j^{(t+1)})(x_i - \hat{\mu}_j^{(t+1)})^T}{\sum_{i=1}^{n} \tau_{ij}^{t}}$$

initialization: e.g. $\mu_j^{(0)}$ from k-mean++

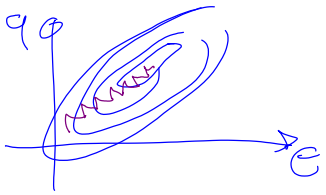
$\Sigma_j^{(0)}$ big spherical variance

$\pi_j^{(0)}$ proportions from k-mean++

$\pi_j^{(0)}$ proportions from k-means++

- one problem with EM:

  sometimes make very slow progress



→ where 'conjugate gradient' method can win

Aside:

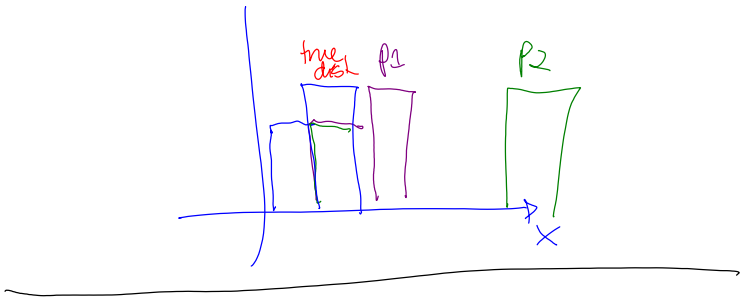
- with ∞-data, true MLE gives the correct parameters

  but it's intractable to find them (non-convex)

- instead, use method of moments to recover parameters "cheaply" with guarantees?!?

  see. [Hsu & Kakade TCS 2013]

https://arxiv.org/abs/1206.5766



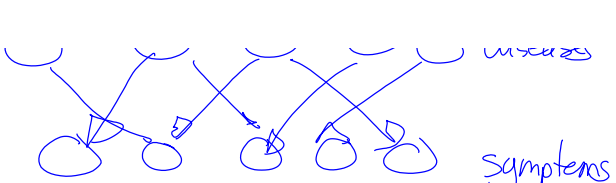
Graphical model

graph model ⇝ probability + C.S.

R.V.s ⇝ graphs

graph ⇝ efficient data structure

GMR



eg. $X_1, \dots, X_n$ R.V.'s

$X_i \in \{0,1\}$ $n \to 100$s

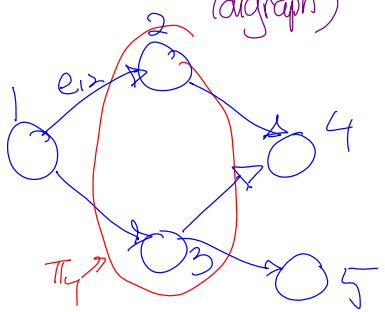
$\Rightarrow 2^{100}$ #'s table $\approx$ untractable

Graph theory

directed graph $G = (V, E)$ (digraph)

$V = \{1, \dots, n\}$ "nodes/vertices"

$E \subseteq V \times V$ "directed edges"



$e_{12} = (1,2)$

directed path $1 \rightsquigarrow 4$

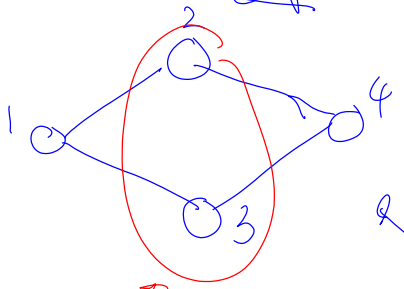
$(1,2), (2,4)$

$\pi_i \triangleq \{j \in V : \exists (j,i) \in E\}$ = set of parents of i

undirected graph: elements of E are 2-sets (note: no self loops)

thus we have $\{i,j\} = \{j,i\}$

vs. $(i,j) \neq (j,i)$ [order matters]

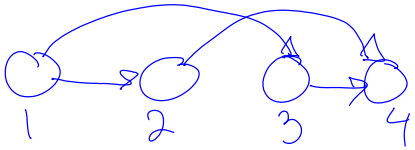


here we have (undirected) path from 2 to 3

neighbors replace the parents/children terminology from digraphs

def: DAG = directed acyclic graph = digraph with no cycle

def: an ordering $I: V \to \{1,\dots,n\}$ is said topological for G
iff nodes in π_i appear before i in I $\forall i$
(i.e. $j \in \pi_i \Rightarrow I(j) < I(i)$)



→ if topological ordering
⇒ all edges go from left to right
["no back edge"]

proposition: digraph G is a DAG $\Leftrightarrow$ $\exists$ topological ordering of G

proof: $\Leftarrow$) trivial: no back edge ⇒ no cycle

$\Rightarrow$) use DFS algorithm to construct a topological sorting in $O(|E|+|V|)$

---

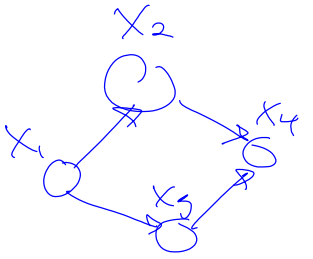
general issues in this class

A) representation → DGM
→ UGM
parameterization → exponential family
} probabilities

B) inference computing $p(x_Q \mid x_E)$
"query" "evidence"

↳ sum-product algorithm

C) statistical estimation → MLE
maximum entropy
method of moments



Notation:

n discrete R.V. $X_1, \dots, X_n$ V = set of vertices one R.V. per node

joint $p(X_1=x_1, X_2=x_2, \dots, X_n=x_n) \overset{\text{shorthand}}{=} p(x_1, \dots, x_n)$
$= p(x_V) \overset{(V)}{=} p(x)$

recall $p(x_{1:n}) = p(x)$

for any $A \subseteq V$

$p(x_A) = P\{X_i = x_i : i \in A\} = \sum_{x_{A^C}} p(x_A, x_{A^C})$

$p(x_A)$: marginal on x_A; subset of "subscripts"

x_{A^C}: summing over all possible values of $\{x_i\}_{i \in V \setminus A}$

$x_{\{1,2,4\}}$ or $\{x_1, x_2, x_4\}$ $A^C = V \setminus A$ $\{x_i\}_{i \in V \setminus A}$

revisit conditional independence:

let $A, B, C \subseteq V$ $X_A \perp\!\!\!\perp X_B \mid X_C$

(F) $\iff p(x_A, x_B \mid x_C) = p(x_A \mid x_C)\, p(x_B \mid x_C)$ $\forall x_A, x_B, x_C$ s.t. $p(x_C) > 0$

(C) $\iff p(x_A \mid x_B, x_C) = p(x_A \mid x_C)$ (for x_B, x_C s.t. $p(x_B, x_C) > 0$)

"marginal independence" : $X_A \perp\!\!\!\perp X_B \mid \emptyset$

3 facts about C.I.:

1) can repeat variables e.g. $X \perp\!\!\!\perp Y, Z \mid Z, W$ is fine to say

2) decomposition: $X \perp\!\!\!\perp (Y, Z) \mid W \Rightarrow X \perp\!\!\!\perp Y \mid W$ and $X \perp\!\!\!\perp Z \mid W$

3) trick: extra conditioning on both sides of equation maintain true statement

eg $p(x, y) = p(x|y)\,p(y)$ (always true)

$\Rightarrow p(x, y|z) = p(x|y, z)\,p(y|z)$ || ||

(*) pairwise independence $\not\Rightarrow$ mutual independence

e.g. $X_3 = X_1$ xor X_2 with $X_1 \perp\!\!\!\perp X_2$ ~ Bernoulli($\frac{1}{2}$)

here $X_1 \perp\!\!\!\perp X_2$, $X_2 \perp\!\!\!\perp X_3$, $X_1 \perp\!\!\!\perp X_3$

but $X_1 \not\perp\!\!\!\perp (X_2, X_3)$

chain rule (always true): $p(x_V) = \prod_{i=1}^{n} p(x_i \mid x_{1:i-1})$

last conditional is $p(x_n \mid x_{1:n-1})$ ← table with 2^n entries

directed graph model $p(x_V) = \prod_{i=1}^{n} p(x_i | x_{\pi_i})$

π_i: parents of i in graph G

$\rightarrow$ tables of $2^{\max|\pi_i|+1}$