Lecture 11 - EM & GMM

October 8, 2021 3:02 PM

today : finish EM algorithm + GMM MLE
graph theory & DGM

properties of EM algorithm

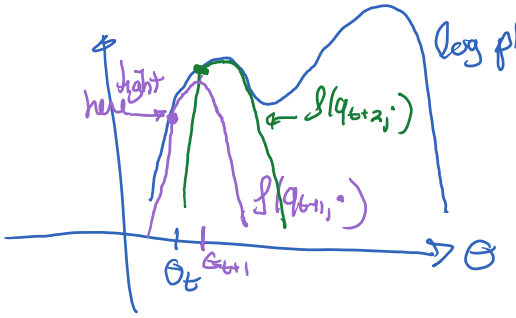
recall : block coordinate ascent on $\mathcal{L}(q, \theta) \leq \log p(x;\theta) \; \forall \theta \& q$

note : $\log p(x;\theta_t) = \mathcal{L}(q_{t+1}, \theta_t)$

↳ where $q_{t+1}(z) = p(z|x;\theta_t)$

properties :

a) $\log p(x;\theta_{t+1}) \geq \log p(x;\theta_t)$ proof : $\log p(x;\theta_{t+1}) \geq \mathcal{L}(q_{t+1}, \theta_{t+1}) \geq \mathcal{L}(q_{t+1}, \theta_t) = \log p(x;\theta_t)$ ($\geq \mathcal{L}(q_{t+1}, \theta) \; \forall \theta$)



b) θ_t in EM converges to a stationary pt. of $\log p(x;\theta)$

ie. $\nabla_\theta \log p(x;\theta)\big|_{\hat{\theta}} = 0$

like k-means, initialization is crucial
→ usually do random restarts

for GMM
could use k-mean++ to initialize the μ's

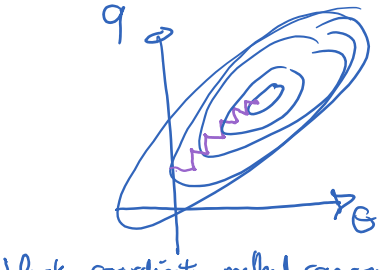
c) $\mathcal{L}(q,\theta) = \mathbb{E}_q\left[\log \frac{p(x,z;\theta)}{q(z)}\right]$

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

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

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

we will revisit this for variational inference $q \in Q$



block-coordinate method can sometimes be slow

For GMM model:

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

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

shorthand to say $z_{i,j} = 1$

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

notation: $x = x_{1:n}$

$z = z_{1:n}$

exercise:

$$p(z|x) = \prod_i p(z_i|x) = \prod_i p(z_i|x_i)$$

complete log-likelihood:

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

($\log p(x_i|z_i;\theta)$: Gaussian; $\log p(z_i;\theta)$: multinoulli)

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

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

$E_q[\mathbb{1}\{z_{ij}=1\}] = q(z_{ij}=1)$

$E_q[z_{ij}] = q(z_{ij}=1)$ [marginal dist.]

during EM, $q_{t+1}(z) = p(z|x;\theta_t)$

weight $\tau_{ij}^t \triangleq p(z_{ij}=1|x;\theta_t) = q_{t+1}(z_{ij}=1)$

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

$$= \prod_i p(z_i|x_i;\theta_t)$$

$$\Rightarrow q_{t+1}(z_i) = p(z_i|x_i;\theta_t)$$

$$\propto p(x_i|z_i;\theta_t)\, p(z_i;\theta_t)$$

($p(x_i|z_i;\theta_t)$: Gaussian; $p(z_i;\theta_t)$: $\pi_{z_i}^t$)

$$\tau_{ij}^t = q_{t+1}(z_{ij}=1) = \frac{\pi_j^{(t)} N(x_i|\mu_j^{(t)},\Sigma_j^{(t)})}{\sum_{\ell=1}^K \pi_\ell^{(t)} N(x_i|\mu_\ell^{(t)},\Sigma_\ell^{(t)})} \quad \left\{ p(x_i, z_{ij}=1|\theta^{(t)}) \right\} \left\{ p(x_i|\theta^{(t)}) \right\}$$

E step for GMM: computing $\tau_{ij}^{(t)}$ for $i=1,\dots,n$ using $\theta^{(t)}$

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

exercise:

$$\hat{\pi}_j^{(t+1)} = \sum_i \tau_{ij}^{(t)}$$ ("soft-counts")

exercise :

M step for EM for GMM

$$\hat{\pi}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{i,j}^{(t)}}{n} \quad \text{"soft-counts"}$$

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

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

- initialize : e.g. $\mu_j^{(0)}$ from kmeans++
  $\Sigma_j^{(0)}$ big spherical covariance $\Sigma_j^{(0)} = \sigma^2 I$ (σ^2 big)
  $\pi_j^{(0)}$: proportions from k-means++

- EM step in GMM with fixed $\Sigma_j = \sigma^2 I$ with $\sigma^2 \to 0$
  ⇝ get k-means algo ?

16h03

Graphical model

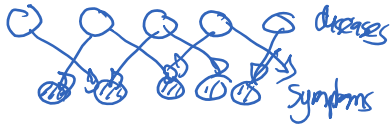
graph. model ⇝ prob theory + C.S.
↓ R.V. ↓ graph

graph → efficient data structure

e.g. $X_1, \dots, X_n$ R.V.
$X_i \in \{0,1\}$

$n = 100$s
⇒ 2^{100} #'s table ⇝ intractable

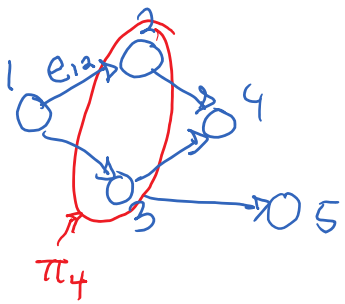
QMR



Graph theory review :

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

$V = \{1, \dots, n\}$ "nodes/vertices"
$E \subseteq V \times V$ "directed edges"



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

directed path $1 \rightsquigarrow 4$ compatible seq. of edges
$(1,2), (2,4)$

π_4

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

set of parents of i

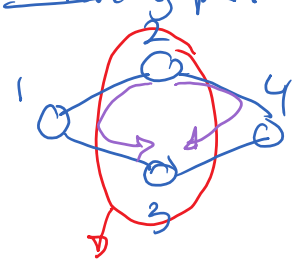
$\text{children}(i) \triangleq \{j \in V : \exists (i,j) \in E\}$

seq. of edges

$(1,2), (2,4)$
or
$(1,3), (3,4)$

undirected graph: $G = (V, E)$ where elements of E are 2-sets (set of 2 elements)

(note: no self loop in this class ie. $|e|=2$)



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

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

undirected path $2 \rightsquigarrow 3$

"neighbors" of node 4 or 1

$N(i)$ for neighbors of i

$= \{j \in V : \exists \{i,j\} \in E\}$

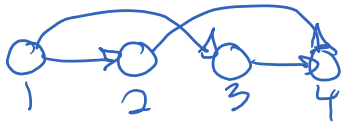
neighbors replace the parents/children terminology from a digraph

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

def: an ordering $I : V \to \{1, \dots, |V|\}$ is said to be topological for digraph G

iff nodes in π_i appear before i in I $\forall i$ (n)

$(j \in \pi_i \Rightarrow I(j) < I(i))$

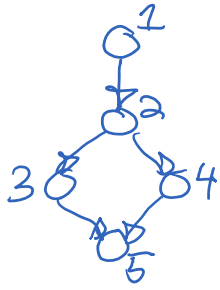


→ if top ordering $\Rightarrow$ all edges go from left to right [no "back edge"]

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

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

$\Rightarrow$ use DFS algorithm to label nodes in decreasing order when have no children



to construct a top. sort.

in $O(|E| + |V|)$

top. sort → finding a topological ordering

Notation for graph models

n discrete R.V. $X_1, \dots, X_n$

↳ discrete R.V. for simplicity conditional dist. concept is

↳ discrete R.V. for simplicity

conditional dist. concept is tricky to formalize for cts. R.V. (see Borel-Kolmogorov paradox)

V ← set of vertices, one R.V. variable per node

joint $p(X_1 = x_1, \dots, X_n = x_n) \overset{\text{shorthand}}{=} p(x_1, \dots, x_n)$ $\quad x = x_{1:n}$

$= p(x_V) \overset{(P)}{=} p(x)$

for any $A \subseteq V$

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

x_A: subset of "subscripts"

x_{A^c}: summing over all possible values of x_{A^c} can take, i.e. $(x_i)_{i \in A^c = V \setminus A}$

question: is $p(x_1, x_2, x_4) \overset{?}{=} p(x_2, x_1, x_4)$?

yes? usually in this class, we use "typed convention"

revisit cond. indep.:

let $A, B, C \subseteq V$

(*) $X_A \perp\!\!\!\perp X_B \mid X_C$

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

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

(*) "marginal independence": $X_A \perp\!\!\!\perp X_B \mid (\emptyset \text{ or } X_\emptyset)$

$p(x_A, x_B) = p(x_A)\, p(x_B)$

2 facts about cond. indep.:

1) can repeat variables in statement (for convenience) $X \perp\!\!\!\perp Y, Z \mid Z, W$ is equivalent to $X \perp\!\!\!\perp Y, Z \mid W$ is fine to say

(does not do anything)

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

(*) pairwise indep. $\not\Rightarrow$ mutual indep → see lecture 3

$Z = X \oplus Y$ (XOR)

(*) chain rule (always true) $p(x_V) = \prod_{i=1}^{n} p(x_i \mid x_{1:i-1})$ $\quad$ last cond. $p(x_n \mid x_{1:n-1})$ ← table with 2^n entries

(always true) $\bar{I} = 1$

assumption in DGM:

$p(x_V) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})$ → tables of $2^{\max |\pi_i| + 1}$

$X_i \perp\!\!\!\perp X_{1:i-1} \mid X_{\pi_i}$