Lecture 11 - scribbles

Friday, October 11, 2019 13:58

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

properties of EM algorithm

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

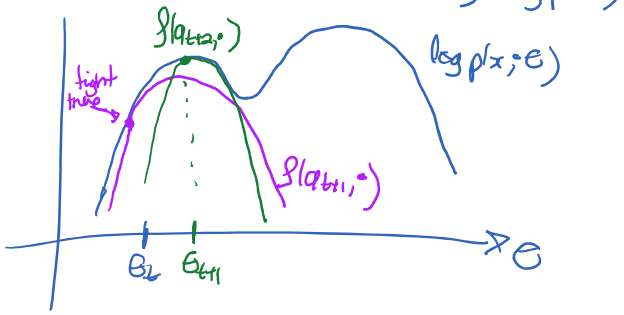
note: $\log p(x;\theta_t) = \mathcal{L}(q_{t+1}, \theta_t)$

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

properties:

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

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



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 use random restarts

for GMM, could use k-means to initialize the μ_k's

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

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

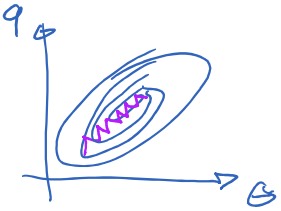
$= \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x,\theta)}\right] \triangleq KL(q(\cdot) \,\|\, p(\cdot|x;\theta))$

KL divergence

(we will revisit this)

for variational inference $q \in \mathcal{Q}$

block-coordinate method can be slow



for GMM model:

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

$x_i \mid z_{ij}=1 \sim N(\mu_j, \Sigma_j)$ ← shorthand to say $z_{ij}=1$

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

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

$z = z_{1:n}$

complete log-likelihood:

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

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

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

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

$\mathbb{E}_q[z_{ij}] = q(z_{ij}=1)$ [marginal distribution]

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

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

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

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

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

($p(x_i \mid z_i;\theta_t)$: Gaussian; $p(z_i;\theta_t)$: π_{z_i})

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

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

$j = 1, \dots, k$

M step : $\max_{\{\mu_j, \Sigma_j, \pi_j\}} \sum_i \sum_j \gamma_{ij}^t [\log p(x_i | \mu_j, \Sigma_j) + \log \pi_j]$

M step for EM for GMM

Exercise :

$$\hat{\pi}_j^{(t+1)} = \frac{\sum_i \gamma_{ij}^t}{n}$$ "soft-count"

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^n \gamma_{ij}^t x_i}{\sum_{i=1}^n \gamma_{ij}^t}$$ soft-cluster assignment

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

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

$\Sigma_j^{(0)}$ big spherical variance $\Sigma_j^{(0)} = \sigma^2 I$, σ^2 big

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

EM step in GMM with fixed $\Sigma_j = \sigma^2 I$ with $\sigma \to 0$

$\Leftrightarrow$ k-means

14h55

Graphical model

graph model $\rightsquigarrow$ prob. theory + C.S.

R.V. — graph

graph $\rightsquigarrow$ efficient data structure

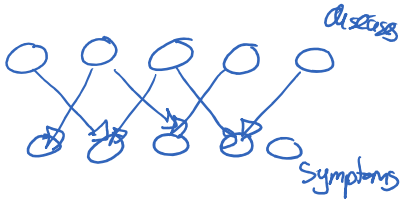
e.g. $X_1, \dots, X_n$ R.V.'s

$X_i \in \{0, 1\}$

$n = 100$

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

QMR

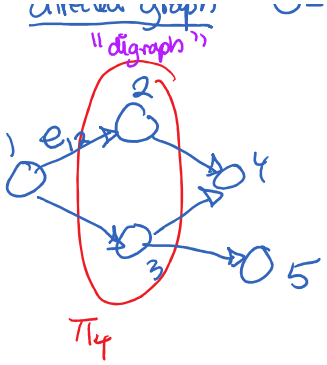


Graph theory review:

directed graph "digraph" $G = (V, E)$ $V = \{1, \dots, n\}$ "nodes/vertices"

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

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



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

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

directed path $1 \rightsquigarrow 4$ → compatible sequence of edges

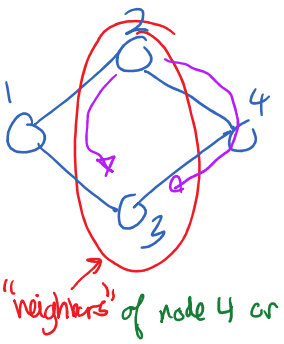
$(1,2), (2,4)$ or $(1,3), (3,4)$

π_4

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

set of parents of i

undirected graph: $G = (V, E)$ elements of E are 2-sets (set of 2 elements) Note: no self loop $\Leftrightarrow |e| = 2$



"neighbors" of node 4 or 1

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

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

undirected path $2 \rightsquigarrow 3$

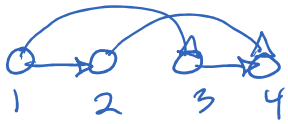
(*) neighbors replace the parents/children terminology from digraphs

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

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

iff nodes in π_i appear before i in I $\quad \forall i$

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



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

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 construct a top. sort in $O(|E|+|V|)$

Notation for graph models:

n discrete R.V. $X_1, \dots, X_n$ → discrete R.V. for simplicity

$V \leftarrow$ set of vertices

conditional dist. concepts tricky to formalize

see Borel-Kolmogorov paradox

one R.V. per node shorthand

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

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

for any $A \subseteq V$

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

A: subset of "subscripts"

$x_{\{1,2,4\}} \leftrightarrow \{x_1, x_2, x_4\}$

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

Question: is $p(x_1, x_2, x_4) = p(x_2, x_1, x_4)$?

yes usually in this class ("typed convention")

Revisit cond. indep.:

Let $A, B, C \subseteq V$

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

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

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

(*) "marginal indep.": $X_A \perp\!\!\!\perp X_B \mid \emptyset$ ($X_\emptyset$)

2 facts about C.I.:

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

Z: does not do anything?

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

$X \perp\!\!\!\perp Z \mid W$

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

$Z = X \oplus Y$ (xor)

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

assumption in DGM

a last cond. $p(x_n \mid x_{1:n-1})$ / table with $\sim 2^n$ entries

assumption in DGM

$= \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})$ → tables of $2^{\max_i |\pi_i|+1}$ entries ($\sim 2^{r}$ entries)

Directed graphical model

Let $G = (V, E)$ be a DAG

a **directed graphical model (DGM)** (associated with G) (aka Bayesian network)

is a **family** of distributions over X_V

where $\mathcal{L}(G) = \{ p \text{ is a dist. over } X_V : \exists \text{ legal (not nec. unique) factors } f_i \text{ s.t.} \}$

$$p(x_V) = \prod_{i=1}^{n} f_i(x_i \mid x_{\pi_i})$$

(potential sets) $f_i : \Omega_{X_i} \times \Omega_{X_{\pi_i}} \to [0, 1]$

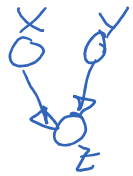
s.t. $\forall i \sum_{x_i} f_i(x_i \mid x_{\pi_i}) = 1 \quad \forall x_{\pi_i}$

f_i is like a CPT (conditional prob table)

terminology: if we can write $p(x_V) = \prod_i f_i(x_i \mid x_{\pi_i})$ (this determines G)

then we say that "p factorizes according to G"

one example:



$p \in \mathcal{L}(G) \iff p(x, y, z) = f_x(x) f_y(y) f_z(z \mid x, y)$

very soon, we show that $p(x_i \mid x_{\pi_i}) = f_i(x_i \mid x_{\pi_i})$