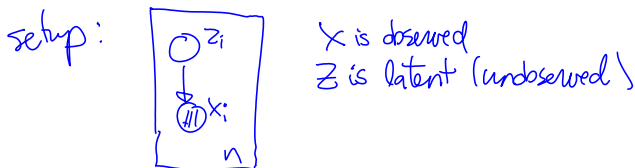


# Lecture 9 - scribbles

Friday, September 30, 2016  
13:26

today: EM & GMM

EM - max likelihood in latent variable model



$$\begin{aligned} \text{log-likelihood } \log p(x; \theta) &= \log \prod_{i=1}^n p(x_i; \theta) \\ &= \sum_{i=1}^n \log p(x_i; \theta) \\ &= \sum_{i=1}^n \log \sum_{z_i} p(x_i, z_i; \theta) \end{aligned}$$

→ gives multi-modal difficult optimization

options for ML in latent variable model

- 1) do gradient ascent or conjugate gradient
- 2) EM algorithm → coordinate ascent on auxiliary function that lower bounds  $\log p(x; \theta)$

nice interpretation in terms of filling in missing data

$$\log \sum_z p(x, z)$$

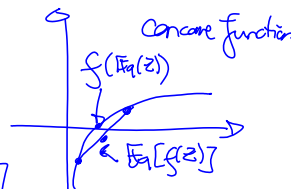
$$\log \sum_z q(z) \frac{p(x, z)}{q(z)}$$

where  $q(z)$  is  
some distribution on  $z$

Jensen's inequality:

$$\mathbb{E}_q[f(z)] \leq f(\mathbb{E}_q(z))$$

when  $f$  is concave



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

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

Jensen's inequality


trick!

$$\leq \mathcal{J}(q, \theta) = \mathbb{E}_{q(z)} [\log p(x, z; \theta)] + H(q)$$

auxiliary function

$$\log p(x; \theta) \geq \mathcal{J}(q, \theta) \quad \forall q, \theta$$

$$\log p(x; \theta) \geq J(q, \theta) \quad \forall q, \theta$$

we get equality when  $\frac{p(x, z)}{q(z)} = \text{constant with respect to } z = p(x)$  

i.e.  $q^*(z) \propto p(x, z)$

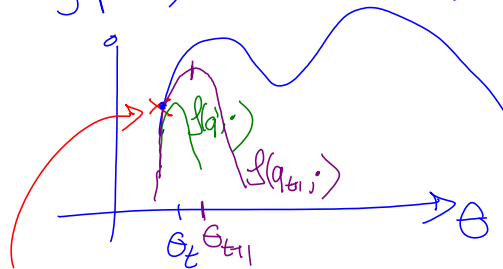
$$\Rightarrow q^*(z) = \frac{p(x, z)}{\sum_z p(x, z)} = \frac{p(x, z)}{p(x)} = p(z|x)$$

this means that  $\arg \max_{q \in \text{distributions}} J(q, \theta_t) = p(z|x; \theta_t)$

EM algorithm: E step:  $q_{t+1} = \arg \max_q J(q, \theta_t) \rightarrow q_{t+1}(z) \triangleq p(z|x; \theta_{t+1})$

M step:  $\theta_{t+1} = \arg \max_{\theta} J(q_{t+1}, \theta)$

coordinate ascent on  $J(q, \theta) \leq \log p(x; \theta)$  



we have  $\log p(x; \theta_t) = J(q_{t+1}, \theta_t)$   
 $\hookrightarrow p(z|x; \theta_t)$

properties: a)  $\log p(x; \theta_{t+1}) \geq \log p(x; \theta_t)$  [ascent alg on  $\theta$ ]

b) converges to stationary pt. of  $\log p(x; \theta)$  [ $\nabla_{\theta} \log(\cdot) = 0$ ]

but not global max in general

like k-means, initialization is crucial

$\rightarrow$  random restarts

• example could use k-means++ to initialize GMM

Operationally =

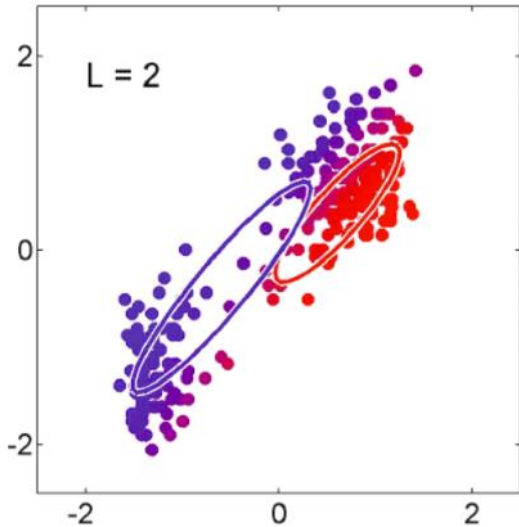
E step:  $q_{t+1} = \arg \max_q J(q, \theta_t) \rightarrow q_{t+1}(z) = p(z|x; \theta_t)$   
 [inference]

M step:  $\theta_{t+1} = \arg \max_{\theta} J(q_{t+1}, \theta)$

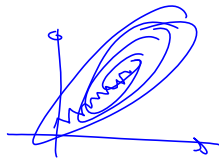


$\Sigma_j$  big spherical variance

$\pi_j$  proportion from K-means



one problem with EM: (sometimes) slow progress



→ where conjugate gradient can win

Aside: with  $n$ -data, true ML gives correct parameters, but intractable to find [non-convex]

• instead, use method of moments to recover parameters with guarantees?

(see e.g. [Hsu & Kakade TCS 2013])

<https://arxiv.org/abs/1206.5766>

## Graphical model (??)

graph model → probability + CS.

R.V. ↔ graphs

QMR

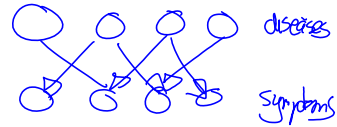
Graph → efficient data structure

example:



Graph  $\rightarrow$  efficient data structure

example:



$X_1, \dots, X_n$  random variables

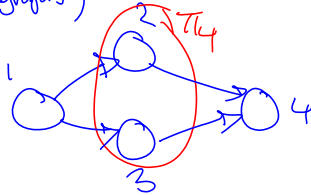
$X_i \in \{0,1\}$

$n \rightarrow 1005$

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

Graph theory:

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

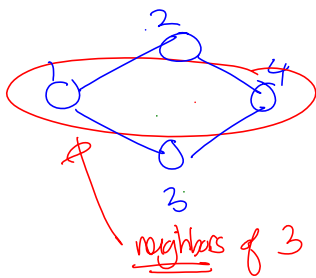


$E \subseteq V \times V$  "directed edges"

directed path  $1 \rightarrow 2 \rightarrow 4$   
 $(1,2), (2,4)$

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

undirected graph: elements of  $E$  are 2-sets



$\{i,j\} = \{j,i\}$

note:  
(no self loops)

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

$\rightarrow$  path from 2 to 3 here

neighbors replace the parents & children terminology from directed graphs

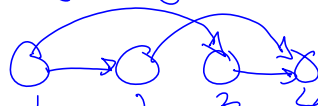
def: DAG = directed acyclic graph  $\rightarrow$  directed graph with no cycle

an ordering  $I: V \rightarrow \{1, \dots, n\}$  is topological for G

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

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

$\rightarrow$  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

⇒) Use DFS to construct top sort in  $O(|E|+|V|)$

(see [https://en.wikipedia.org/wiki/Topological\\_sorting#Depth-first\\_search](https://en.wikipedia.org/wiki/Topological_sorting#Depth-first_search) )