

16. PGMs: EM Algorithm

MA: focus on PGMs where only subset observed

MLE for general BNs

- (1) write down log-likelihood
- (2) estimate parameters (prob.) of CPT

(3) plate notation

- IID or exchangeable

recognizable likelihood of BN.

(4) write down likelihood of the graph

(5) split into components → à la lecture 5

estimate param. of components

MLE for BNs with tabular CPDs

- parameters are values in CPD
- count instances
- compute sufficient statistics
- compute log-likelihood
- max likelihood to estimate param using constrained opt (prob. constr.)

HMM

- supervised & un-supervised
- supervised → observe <sup>not</sup> dice/states
- unsupervised → latent states (cannot observe what dice)

HMM - supervised

- i) require initial, transmission, emission probabilities
- ii) rows of matrix sum to one → CPD

## Supervised ML estimation

- Have sequence of states AND observations
  - ↳ can count; calculate empirical probabilities
- ML estimates are framed in terms of counts
  - Q: what are subscripts and superscripts inside summations in  
 $a_{ij}^{ML} \rightarrow$  arise from one not encoding

## Intuition

- (\*) captured in slides ✓
- outfitting for small datasets
- see example

- transition probability from fair  $\rightarrow$  unfair (loaded) estimated as 0.
  - over a limited sequence of runs in which only four dice states are observed.
- Q: can you reasonably expect to see this as a deficiency given you haven't observed loaded die states?

## suggested solution: pseudocounts

- Add pseudocounts; which encode prior beliefs  
(to count instances)
  - encodes strength of belief; total no. msg instances
  - Bayesian estimation uses a uniform prior with param. strength = pseudocounts
- (\*) large total pseudocounts  $\rightarrow$  strong prior belief  
smaller total pseudocounts  $\rightarrow$  avoid 0 probabilities (smoothing)

(\*) partially observed GMs

(\*) HMMs for speech recognition prior to deep learning

- latent states from which observations are noisy sampling

(\*) Genomics

- inference as a subroutine for parameter estimation.

(\*) (C) - examine equations here during review

2 standard probabilistic inference tasks

- PGMs require iterative solving

task 1 & task 2

(\*) next module → approximate inference techniques  
(<sup>analytic</sup> computationally infeasible integrals)

mixture models

(\*) Infer as a subroutine for learning - multimodal  $p(x)$

- represent a complex density through mixture of simple ones.

- simpler densities → unimodal/Gaussian/analytically tractable

- write down latent variables (discrete) that select component densities  
(auxiliary)

- after model fitted; every component of density will  $\neq$  select a  
sub-population

(\*) unobserved variables

- z variable allows decoupling of complex distri into simple components

GMMs /  $n^{\text{th}}$  training instance

$$(*) p_{\text{mix}}(\mu, \Sigma) = \sum_R \pi_R N(x | \mu_R, \Sigma_R) \quad (\text{likelihood of a single point})$$

\underbrace{\hspace{10em}}\_{\substack{\text{comp.} \\ \text{mix pop.}}}

- mixture proportions unknown; ad. parameters of components unknown
- generative model specification

② Why is parameter estimation / learning harder?

- fully observed models  $\rightarrow$  log likelihood decomposes into a sum of local terms

(\*)

- decoupled

- with latent variables;  $\rightarrow$  parameters become coupled together via marginalisation

(JP presented this as log, sum, product difficulties)

(\*) Distinction: dependent variables; IID samples (?)

Façade EM

- we're assuming completely observed data (as a reference for unobserved case)

- (1) (2) (3): review (should be very familiar)

K-means

- latent variable indicating data point cluster

- assign most probable cluster to data point  
(<sup>cluster</sup> closest centroid to data point)

- init clusters

- assign data point to cluster } Heate  
- recompute centroids } to convergence

## GMM (Example)

- expected complete log-likelihood

$$\begin{aligned} \langle \ell_c(\theta; x, z) \rangle &= \sum_n \langle \log p(z_n | \pi) \rangle_{p(z|x)} + \sum_n \langle \log p(x_n | z_n, \mu, \Sigma) \rangle_{p(z|x)} \\ &= \sum_n \sum_K \langle z_n^K \rangle \log \pi_K - \frac{1}{2} \sum_n \sum_K \langle z_n^K \rangle ((x_n - \mu_K)^T \Sigma_K^{-1} (x_n - \mu_K) \\ &\quad + \log |\Sigma_K| + C) \end{aligned}$$

(\*) Take expectation of complete log like;  
wrt the condit. distri of latent given observed (auxiliary post.)

## E-step

- compute expected value of sufficient statistics of hidden variables  
(i.e.  $z$ ) given current est of param

$$z_n^{(t)} = \langle z_n^K \rangle_{q(t)} = p(z_n^K = 1 | x, \mu^{(t)}, \Sigma^{(t)}) = \frac{p(z_n^K = 1, x; \mu^{(t)}, \Sigma^{(t)})}{p(x; \mu^{(t)}, \Sigma^{(t)})}$$

(\*) Essentially doing inference here

## M-step

- compute parameters wrt current results of expected value of hidden variables.

(\*) isomorphic to MLE  $\rightarrow$  latent variables replaced by expectations  
in general, replaced by sufficient statistics.

(W@D): Review this all  $\rightarrow$  essentially revision (you know it well  
by now)

EM-GMM is soft version of K-means

## Theory underlying EM

- if not observe  $z$ ; computing complete log like is diff:-

$$\ell_c(\theta; D) = \log \sum_z p(x, z | \theta) = \log \sum_z p(z | \theta) p(x | z, \theta)$$

3 ingredients (to better understand)

- complete log-likelihood

- incomplete log-likelihood

- expected complete log-like. (with a q-distribution over  $z$ s, condition on  $x$  and  $\theta$ )

MR: If we maximise expected complete-log-like  
over q-distr; what happens to our objective  
(incomplete/marginal log like)

MR: Apply Jensen's inequality to get a bound on the  
log marginal likelihood (which one specified in terms  
of expectation of complete log-like)

$$\begin{aligned}
 (*) \quad \ell(\theta; x) &= \log p(x | \theta) = \log \sum_z p(x, z | \theta) = \log \sum_z q(z | x) \frac{p(x, z | \theta)}{q(z | x)} \\
 &\geq \sum_z q(z) \log \frac{p(x, z | \theta)}{q(z | x)}
 \end{aligned}$$

$$\Rightarrow \ell(\theta; x) \geq \langle \ell_c(\theta; x, z) \rangle_q + H_q$$

$\underbrace{\hspace{1cm}}$  lower bound

When is lower  
bound tight?

- for fixed data  $x$ ; define a functional called the free energy:-

$$f(q; \theta) \stackrel{\text{def}}{=} \sum_z q(z | x) \log \frac{p(x, z | \theta)}{q(z | x)} \leq \ell(\theta; x)$$

(\*) EM algorithm is co-ordinate ascent on  $F$ :

$$\text{E-step: } q^{t+1} = \underset{q}{\operatorname{argmax}} F(q, \theta^{(t)})$$

$$\text{M-step: } \theta^{t+1} = \underset{\theta}{\operatorname{argmax}} F(q^{t+1}, \theta^{(t)})$$

- maximise functional wrt  $q$ -distri (E-step)
  - fix  $q$  and optimise free energy with respect to param (M-step)
- (\*) EM is special case of MM  
(co-ordinate ascent on  $F$ )
- repeat all possible distri  $q$ s on  $x$  axis; and all possible param  $\theta$  on  $y$ -axis.

### E-step (analysis)

$$q^{t+1} = \underset{q}{\operatorname{argmax}} F(q, \theta^{(t)}) = p(z|x, \theta^{(t)})$$

(\*) posterior distri over latent variables given data; and param.

Proof: setting attains the bound  $I(\theta; z) \geq F(q, \theta)$

Ex: Q: quick check of derivation.

(\*) Alternatives to Jensen :- Use variational calculus or

$$I(\theta; z) - F(q, \theta) = \text{KL}(q || p(z|x, \theta)) \quad (\text{KL div. relationship})$$

obj      free  
energy      KL-div.

-  $\text{KL} = 0$  when  $q = p(z|x, \theta)$  (I.P.)

## M-step

- optimise free negy wrt param.

(\*) free negy breaks into:-

expected complete and entropy term.

log like.

(negy)

(does not depend  
on  $\theta$  is entropy)

## Q&As: review

- GMM (Gaussian Mixture) - simple PGHMS Q: How does this (EM)  
relate to graphical  
models

- insert plate:

### GMM exp

- distil general graphical models  $\rightarrow$  building blocks

- HMM - use of EM (but forward-backward)

- static  $\rightarrow$  dynamic mixture models

- Baum Welch (EM for HMM)

- same strategy as EM

i) write down complete log-like.

ii) write down expected complete log-like. with respect to posterior

- iterate E-step (compute exp) M-step (update param)

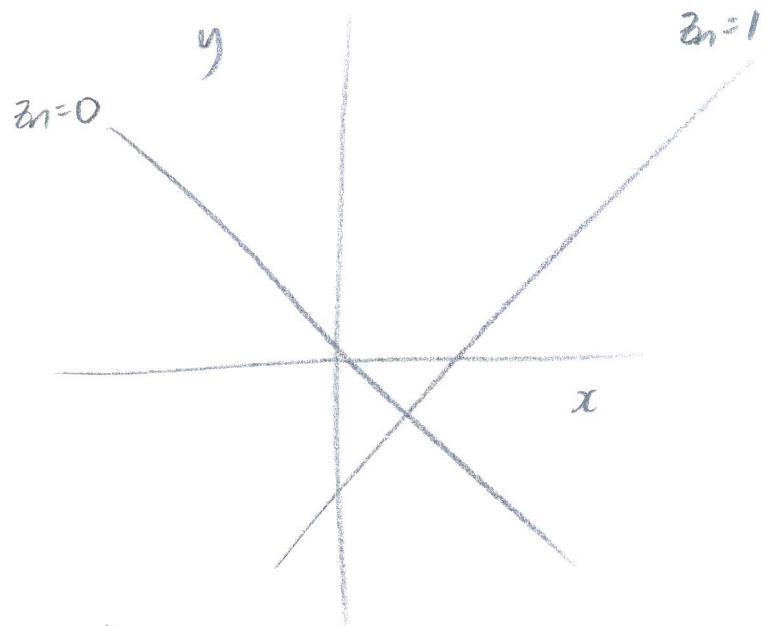
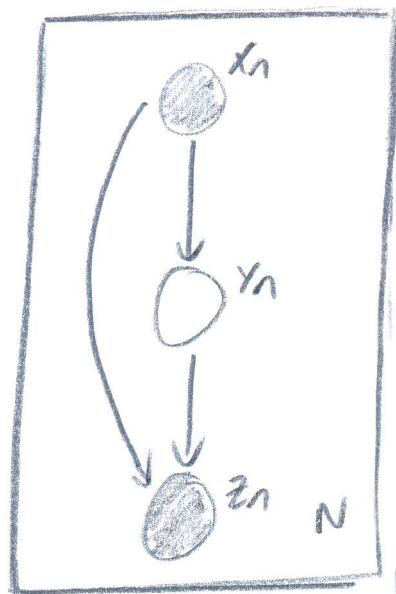
- in general GMM; ~~can~~ use structure to help with E-step

(\*) Summarise EM algorithm so you can remember

## (\*) conditional mixture models: Mix of experts (EM app.)

- ① ② - check you can reconcile plate nemo; digram.
- (\*) tries to model rel. between features  $x$  and response  $y$ ; related through latent  $z$ .

## (\*) relation between $x$ and $y$ - piecewise linear



- ① ② check this model (Murphy / readings)

## (\*) optimise conditional prob.

### GM for conditional mixture

## (\*) IRLS → covered in lecture 5

MIL: <sup>on \*  
complex, rich models if we have prior knowledge</sup> mixture of overlapping experts

- voting
- (stumped our)

## Partially hidden data

- missing variables on some cases; and not on others
- ML: use a mix of ML and EM over a partitioned cost/complete log-likelihood.

## EM variants

- (\*) sparse EM
- (\*) incomplete EM

⑦ Review/fill in notes  
(contd.) ↓