10-708 11/05/2020

L6 - PGMs: EM Algorithm

MA: Focus on PGMs where only subset observed

MLE for general BNs
- (*) write down log-likelihood
(*) estimate parameters (prob.) of CPT

(*) Plate notation
- IID or exchangeable

Decomposable likelihood of BN.
(*) write down likelihood of the graph
(*) 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 & unsupervised
- supervised → observe what dice/states
- unsupervised → latent states (cannot observe what dice)

HMM - supervised
i) Require initial, transmission, emission probabilities
- 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} → arise from one hot encoding

---

- Intuition
- (*) captured in slides ✓
- overfitting for small data sets
- see example
- transition probability from fair → unfair (loaded) estimated as 0.
- over a limited sequence of runs in which only fair dice states are observed.
- (W): can you reasonably expect to see this as a deficiency; given you haven't observed loaded dice states?

- suggested solution: pseudocounts
- Add pseudocounts; which encode prior beliefs (to count instances)
- encodes strength of belief; total no. imag instances
- Bayesian estimation with a uniform prior with param. strength = pseudocounts
- (*) Large total pseudocounts → strong prior belief
  small total pseudocounts → 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.

(*) (10) (20) - examine equations here during review

2 standard probabilistic inference tasks

- PGMs require iteratively solving

task 1 & task 2

(*) next module → approximate inference techniques
(computationally intractable integrals) analytic

mixture models

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

- Represent a complex density through mixture of simple ones.
- simple densities → unimodal/Gaussian/analytically tractable
- write down latent variables (discrete) that select components densities (auxiliary)
- After model fitted, every component of density will ~~be~~ select a sub-population

(*) unobserved variables

- z variable allows decoupling of complex distn into simple components

GMMs

n^{th} training instance

(*) $p(x_n|\mu,\Sigma) = \sum_k \pi_k N(x|\mu_k,\Sigma_k)$ (likelihood of a single point)

mix pop. comp.

- mixture proportions unknown; and parameters of components unknown
- Generative model specification

Q: Why is parameter estimation/learning harder?

- Fully observed models → log likelihood decomposes into a sum of local terms

(*)

- decoupled
- with latent variables; → parameters become coupled together via marginalisation

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

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

Tuned EM

- MLE assuming completely observed data (as a reference for unobserved case)
- (*) (W) (PR): Review (should be very familiar)

K-means

- latent variable indicating data point cluster
- assign most probable cluster to data point (closest centroid to data point)
- init cluster
  - assign data point to cluster
  - Recompute centroids

} iterate to convergence

GMM (Example)

- Expected complete log-likelihood

$$\langle \ell_c(\theta; x, z) \rangle = \sum_n \langle \log p(z_n|\pi) \rangle_{p(z|x)} + \sum_n \langle \log p(x_n|z_n, \mu, \Sigma) \rangle_{p(z|x)}$$

$$= \sum_n \sum_k \langle z_n^k \rangle \log \pi_k - \frac{1}{2} \sum_n \sum_k \langle z_n^k \rangle \left((x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) + \log|\Sigma_k| + C \right)$$

(*) Take expectation of complete log like; under 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 params

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

(*) Essentially doing inference here

M-step

- compute parameters w/ current results of expected value of hidden variables.

- (*) isomorphic to MLE → latent variables replaced by expectations in general, replaced by sufficient statistics.

(W)(17): Review this all → essentially revision (you know it well by now)

EM-GMM : soft version of K-means

Theory underlying EM

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

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

---

3 ingredients (to bette understco)

- complete log-likelihood
- incomplete log-likelihood
- Expected complete log-like. (with a q-distri over zs, condition on x and param θ)

MR: If we maximise expected complete-log like wrt q-distri, 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)

$$(*)\quad \ell(\theta; \underline{x}) = \log p(\underline{x}|\theta) = \log \sum_{\underline{z}} p(\underline{x}, \underline{z}|\theta) = \log \sum_{\underline{z}} q(\underline{z}|\underline{x}) \frac{p(\underline{x}, \underline{z}|\theta)}{q(\underline{z}|\underline{x})}$$

$$\geq \sum_{\underline{z}} q(\underline{z}) \log \frac{p(\underline{x}, \underline{z}|\theta)}{q(\underline{z}|\underline{x})}$$

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

(lower bound)

When is lower bound tight?

- For fixed data $\underline{x}$; define a functional called the free energy:-

$$F(q; \theta) \overset{\text{def}}{=} \sum_{\underline{z}} q(\underline{z}|\underline{x}) \log \frac{p(\underline{x}, \underline{z}, \theta)}{q(\underline{z}|\underline{x})} \leq \ell(\theta; \underline{x})$$

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

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

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)

- represent all possible distri qs on x axis; and all possible param θ on y-axis.

- E-step (analysis)

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

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

Proof: setting attains the bound $l(\theta; \underline{x}) \geq F(q, \theta)$

(E) (F4): quick check of derivation.

(*) Alternatives to Jensen: - use variational calculus or

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

obj free energy KL-div.

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

M-step

- optimise free engy wrt param.

(*) free energy breaks into:-

expected complete log like. (engy) and entropy term (does not depend on θ is entropy)

(W) (RS) review

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

- insert plate:

GLIMLexp

- distil general graphical models → building blocks

- HMM - use of EM (not forward-backward)

- static → 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 GMs; ~~can~~ use structure to help with E-step.

(*) Summarise EM algorithm so you can remember

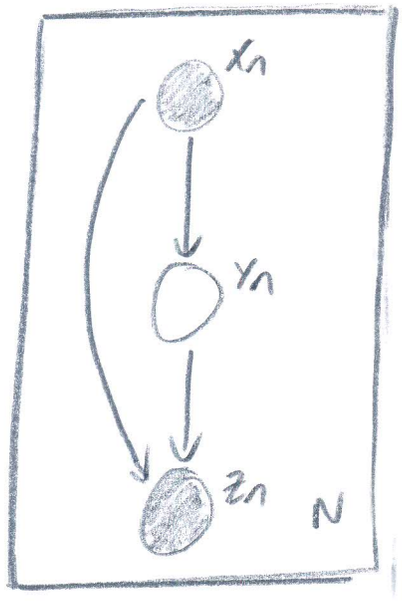
(*) conditional mixture models: Mix of experts (EM app.)

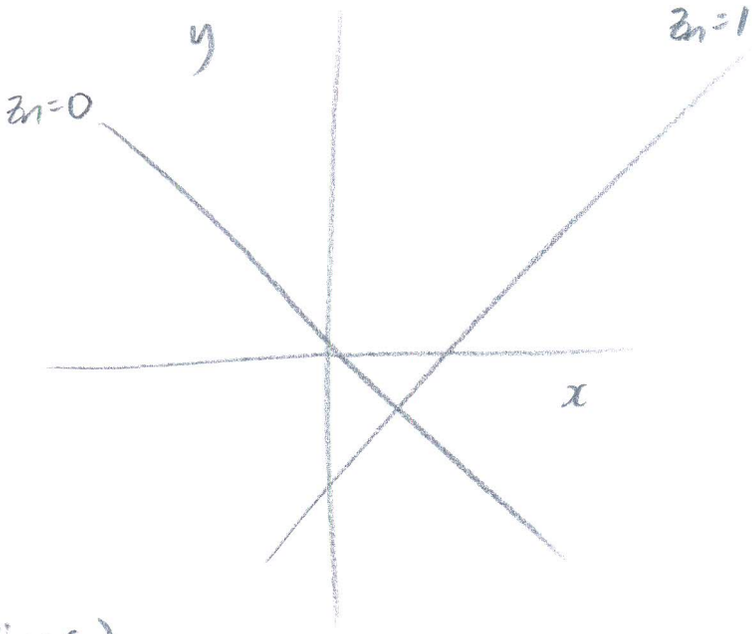
(W)(PRML)

- check you can reconcile plate micro; dicgram.

(*) Tries to model rel. between features x and response y; related through latent z.

(*) relation between x and y - piecewise linear





- (W)(PRML): check this model (Murphy/readings)

(*) optimise conditional prob.

EM for conditional mixture

(*) IRLS → covered in lecture 5

MIL: can use complex, rich models if we have prior knowledge of

---

- mixture of overlapping experts
- voting
- (stacked ove)

Partially hidden data

- Missing variables on some cases; and not on others
- MR: use a mix of ML and EM over a partitioned expected cost/complete log-like.

Ⓦ ⑰ : Review/fill in here (concl.) ↓

EM variants

(*) Sparse EM

(*) Incomplete EM