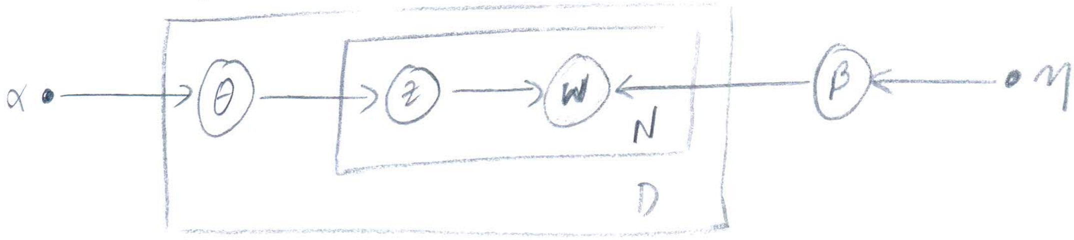
N-708 18/06/2020

UI review - (11b) - Approximate Inference and Topic Models

- Topic models
- Variational inference

(see S.2020 lectures + 10 suppl.)

(*) Generalised topic model - joint likelihood (Topic models / LDA covered in exercises)



- interested in posterior; data likelihood

(*) Inference and learning intractable

$p(\theta_n|D) = ?$ $p(z_{n,m}|D) = ?$ (possible queries)

$p(\theta_n|D) = \frac{p(\theta_n, D)}{p(D)}$ (typo?) - S.2020 → $\frac{p(z_{n,m}|D)}{p(D)}$

$$= \frac{\sum_{\{z_{n,m}\}} \int \left(\prod_n \left(\prod_n p(w_{n,m}|\beta_{z_n}) p(z_{n,m}|\underline{\theta_n}) \right) p(\theta_n|\alpha) \right) p(\beta|\eta) \, d\theta_i \, d\beta}{p(D)}$$

$$p(D) = \sum_{\{z_{n,m}\}} \int \cdots \int \left(\prod_n \left(\prod_n p(x_{n,m}|\beta_{z_n}) p(z_{n,m}|\theta_n) \right) p(\theta_n|x) \right) p(\beta|\eta) \, d\theta_1 \ldots d\theta_N \, d\beta$$

(*) Super exponentially complex. (compute no. terms yourself).

(*) Param estimation → intractable as inference (intractable) cannot be called as a subroutine

(*) V.I.

- Introduce auxiliary/proxy distn q to $p(z)$ (prior in EX slides)
- originally got confused → (q as proxy for true posterior $p(z|x)$)
- But it's the same $p(z|x) = \frac{p(x,z)}{p(x)}$ ← - see that approx. this is indirectly approx. post.

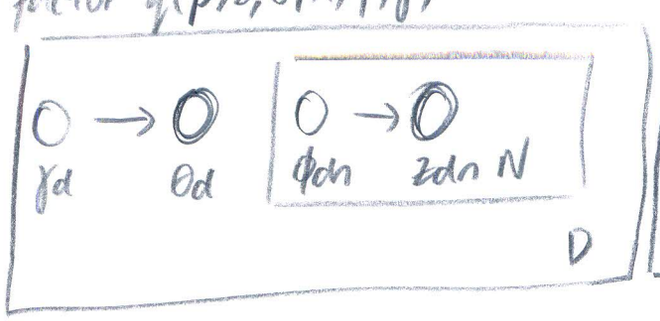
(*) MF Approximation

(*) Parametric form for each marginal factor $q(\beta, z, \theta | \lambda, \phi, \gamma)$

$q(\beta_K | \lambda_K) = \text{Dirich}(\beta_K | \lambda_K)$

$q(\theta_d | \gamma_d) = \text{Dirich}(\theta_d | \gamma_d)$

$q(z_{dn} | \phi_{dn}) = \text{Multi}(z_{dn} | \phi_{dn})$



(*) Estimating param. of variational distn (E-step)

MF Approx

- $q(\beta, \theta, z | \lambda, \gamma, \phi)$

$= \prod_{K=1}^{K} q(\beta_K | \lambda_K) \prod_{d=1}^{D} q(\theta_d | \gamma_d)$

$\prod_{n=1}^{N} q(z_{dn} | \phi_{dn})$

- Aside: - each 'marginal factor' $q(\beta_K), q(\theta_d), q(z_{dn})$ is itself parameterised by variational params (only present in $q(\cdot)$)

v.i. ⟶ i) Find simple $q(\cdot)$ family approx/proxy (fully factored)

ii) Refine parameters within this family.

(*) $\gamma^*, \lambda^*, \phi^* = \underset{\gamma, \lambda, \phi}{\text{argmin}}\ KL(q(\beta, \theta, z | \gamma, \phi) \| p(\beta, \theta, z | w, \alpha, \eta))$

(*) LDA: can compute in closed form.

(*) Some algebraic hard-work now.

(*) Update each marginal.

update - $q(\theta_d) \propto \exp\left\{ \mathbb{E}_{\prod_n q(z_{dn})} \left[\log p(\theta_d | \alpha) + \sum_n \log p(z_{dn} | \theta_d) \right] \right\}$

(i) (ii)

(i) Dirichlet topic proport. distri.

$p(\theta_d | \alpha) \propto \exp\left\{ \sum_{K=1}^{K} (\alpha_K - 1) \log \theta_{dK} \right\}$

(ii) Multinomial word-label frequency distri

$p(z_{dn} | \theta_d) \propto \exp\left\{ \sum_{K=1}^{K} \mathbb{1}(z_{dn} = K) \log \theta_{dK} \right\}$

(*) to obtain:-

$$q(\theta_d) \propto \exp\left\{\sum_{k=1}^{K}\left(\sum_{n=1}^{N} q(z_{dn}=k) + \alpha_k - 1\right)\log\theta_{dk}\right\}$$ ← (OLS1) - get this result.

- Relies on Dirichlet (conjugate prior)

(*) update each msg.

- Similar to $q(\theta_d|\gamma_d)$; obtain optimal param ϕ^*_{dn} for $q(z_{dn}|\phi_{dn})$

$$q(z_{dn}=k|\phi_{dn}) = \phi_{dn}(k) = \beta_{k(w_{dn})}\exp\left\{\psi(\gamma_d(k)) - \psi\left(\sum_{j=1}^{K}\gamma_d(j)\right)\right\}$$

- Reparameterised Multinomial

- And optimal parameters λ^*_k for $q(\beta_k|\lambda_k)$

$$\lambda_k(j) = \eta(j) + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi^*_{dn}\,\mathbb{1}\{w_{dn}=j\}$$

(*) Iterating to convergence yields MF approx. to posterior

(as subroutine?)

- $\lambda_k(j)$ - frequency count of j^{th} word frequency in topic k; which is clear from summation over every doc. and every word and delta fn.

(*) coordinate ascent for LDA

- Have 3 equations for E-step (inference subroutine)

Ex: Key message:-

$p\left(\prod_{j}\right)$ $\quad q\left(\prod_{i}\right) = \prod q_i(y_i) \quad \eta^* = \operatorname{argmin} KL(q\|p)$

(*) variational est. / learning (print S2020 add. slides as suppl.)

- Maximise variational lower bound:-

$$\mathcal{L}(\theta, \phi, \underline{x}) = \mathbb{E}_{q_\phi(\underline{z}|\underline{x})}\left[\log p_\theta(\underline{x}|\underline{z}) + KL(q_\phi(\underline{z}|\underline{x})\|p(\underline{z}))\right]$$

$= \log p(\underline{x}) - KL(q_\phi(\underline{z}|\underline{x}) \| p_\theta(\underline{z}|\underline{x}))$

- previous: =
  - variational E-step
    - maximise $\mathcal{L}$ wrt ϕ with $\underline{\theta}$ fixed
      $\max_\phi \mathcal{L}(\underline{\theta}, \phi; \underline{x})$
  - keep original model params $\underline{\theta}$ fixed; optimise ϕ (variational param.)

- variat. est: -

$$\gamma^*, \lambda^*, \phi^* = \underset{\gamma,\lambda,\phi}{\operatorname{argmin}} KL(q(\beta, \underline{\theta}, \underline{z}|\gamma, \phi) \| p(\beta, \underline{\theta}, \underline{z}|\underline{w}, \alpha, \eta))$$

$\gamma^*, \lambda^*, \phi^*$: variat. params

(*) Additional params in M-step: We want to know $\underline{\theta}$ and β (not present in V.E.) - why?

- computationally not significant (as Bayesian?) (~)
  - $\underline{\theta}$ and β are integrated out

$\alpha \rightarrow \theta \rightarrow z$
(↑ integrated out)

write/specify $p(z|\alpha)$ directly
$Dir(n+\alpha)$

- Hence, no need for $\underline{\theta}$.
- this is why M-step often ignored

- Below is doable,
  - ppl. don't bother

- M-step: estimate $\underline{\theta}$ with ϕ fixed
  $p(\theta|\underline{z}, \underline{x})$
- fix variational param., combine counts.
  $MAP_{\theta,\beta} \mathcal{L}(\underline{\theta}, \phi; \underline{x})$
- or even hyperparam
  $\max_{\underline{\alpha}, \eta} \mathcal{L}(\underline{\theta}, \phi; \underline{x})$

(*) E-step already provides

ϕ_{dn} γ_{dn} λ_K (sufficient statistics for parameters in below) — via exp family; GLMs (?) 0152

↓ ↓ ↓

z_{dn} θ_d β_K

- topical context aug. - clarity

(*) Eg. $\lambda_k(j) = \eta(j) + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi^*_{dn}(k)\, \mathbb{I}(w_{dn} = j)$

- prior pseudocount of that word under any topic
- indic. weighted by n^{th} word being in topic k
- indicator that n^{th} word in document d is "j"!

(*) Estimating $\phi^*, \gamma^*, \lambda^*$ → variational E-step; gives you 'everything else' as they are sufficient statistics

---

(*) Evaluation of ground truth in topic models

- difficult to run on 'synthetic' ground truth
- what is 'ground truth' of a document topics?

(*) V.I. - v. important

MF - approx - vi mp - can 'solve' arbitrarily complex models (vague)
- amounts to removal of edges
- KL between tree od proxy

(*) 0153: see mahomed slides for more on

LRA, VI. (2016) NIPS.