10-708 10/06/2020

L11 - Approximate Inference and Topic Models (Mean Field and Loopy BP) S.2019 11b

- (*) Looks like lecture 11a. VI and loopy B.P. not covered
- (*) Approximate inference methods started as 'tricks', ones which worked then proved.

EX: Maintain 'art of modelling' trajectory, present an example that exposes the concrete need for approx. inference

---

(*) Probabilistic Topic Models

- started as a class project.

(*) How to get started for a new modelling task?

(GT): There is gold here: a way of thinking about the art of modelling

EX: Start with a concrete task/problem you want to solve.
- methods are invoked in service of the task!

e.g. Bird's eye view of 1 million documents.
- Task → clustering; give back cluster label.
  - ↳ embedding → visualisation of cluster labels
- Representation of data → e.g. continuous, binary, counts? (design choice)

(*) Consider each element one at a time

(*) Tasks - document embedding.
- Have each document embedded in a space.

(*) Summarising data using topics
- want the embedding to be meaningful.

(*) See how data changes over time
- Representation of topics may evolve.

(*) user interest topic modelling
- secondary task

(*) Representation
- Bag of words rep.
- convert article → BOW.

(*) For each document, count no. of words, order does not matter.
(*) Each document is a vector in word-space

EX: Benefit and costs? (of representation)
- Benefits → storage, hash tables (makes data simple)
- comparing documents of different lengths (1 x 100 word vs novel
- Probability based on word-orderings → large docs give smaller likelihood (due to product)
  - cannot compare.
- allows comparability.
- costs → ordering may be important for semantics

(*) BOW is a baseline representation.

(*) How to model semantics?
- without using orderings

EX: mental analogy → certain keywords give a higher prob. of a document coming from a topic.
- match a topic with keywords ⊛ intuition.

(*) A topic is a vector of words (in given all vocab.)
A document contains topics in a p (mixing proportion)

(*) ~~A to~~ Note info. compression: a document gets compressed into a weighted sum of topics (low-dimensional with some semantic meaning)

actually a vector of probabilities indicating topic mixing propor

(*) Can then compare similarity of documents

(*) Topic Models - Big Picture

unstructured collection $\xrightarrow{\text{topic discovery}}$ structured topic network

unstructured collection $\xrightarrow{\text{dim. proj.}}$ topic simplex

unstructured collection $\downarrow$ word simplex

word simplex $\xrightarrow{\text{implicit dim. } \downarrow}$ topic simplex

(*) A topic corresponds to a point on the word simplex.
(*) A document — " — on the topic simplex

(*) LSI vs Topic Model (prob LSI)

- likelihood model

$$\underset{\text{words}\times\text{docs}}{\underline{X}} = \underset{\text{words}\times\text{topics}}{\underline{W}}\;\underset{\text{topic}\times\text{topic}}{\underline{\Lambda}}\;\underset{\text{topic}\times\text{documents}}{\underline{D}^T}$$

$$\underset{\text{words}\times\text{docs}}{p(\underline{w})} = \underset{\text{words}\times\text{topics}}{p(\underline{w}|\underline{z})}\;\underset{\text{topics}\times\text{documents}}{p(\underline{z})}$$

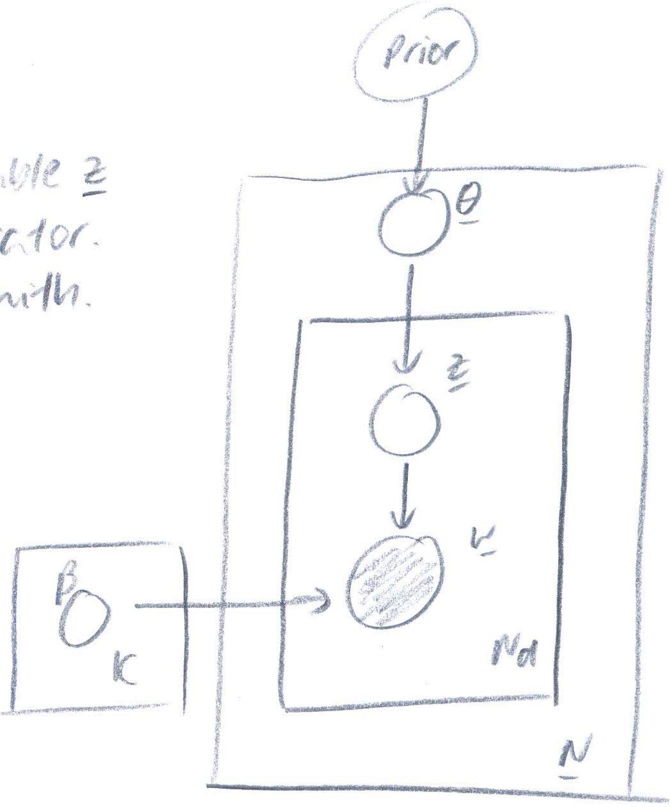
- LSI → word-document matrix decomposition (linear algebraic)
- Topic Models ⟶ conceptually similar (probabilistic inference)
- matrix decomposition techniques (e.g. LSI) make it difficult to bypass computationally expensive matrix inversions (especially for large matrices) (batch)

(*) Probabilistic methods allow ways of dealing with this issue in a piece-wise iterative manner (e.g. LASSO vs batch LR)
i.e. local operations to

(*) Admixture Models

- skip

(*) Topic Models



- Generative process for a document.
  - For every word there is a latent variable z indicating its topic assignment/indicator.
  - i.e. what topic a word is affiliated with.
- Topic indicator z comes from a weight vector θ
- Every document is a vector of topical weights
- Prior: Distn of different topics in corpus.

Generating:

1) Draw a topical weight θ vector/document from a prior distribution

- Given this, for every nth out of Nd words:
  - Draw a topic indicator $z_n \sim \text{multinomial}(\theta)$
  - conditioning on the topic indicator of the nth word z_n, sample the word (using a coll. of word-frequency distns)
  - draw $w_n | z_n, \{\beta_k\}$ from $\text{multinom}(\beta_{z_n})$

(!): Presentation is not the clearest

(*) Choices of priors

- Dirichlet (LDA), Blei et al. 2003
- Logistic Normal, Blei & Lafferty (2005), Ahmed & Xing (2006).

(see the facets of these wrt modelling intuition)

(*) Generative semantics of CoNTAM

- captures intuition of topics being highly correlated
e.g. sports and topics
- use a covariance matrix?
  - only in MVGs (at simplest).
- ok, so use MVGs $\longrightarrow$ to yield γ

ISSUE: we cannot use γ to sample z $\rightarrow$ γ is Gaussian vector with -ve components; NOT Multinomial

- Apply a transformation, apply exponential, then normalise to get appropriate prior.
  $\gamma \rightarrow \theta$ (logistic normal)

EX: Model design can be arbitrary, flexible, to a degree.

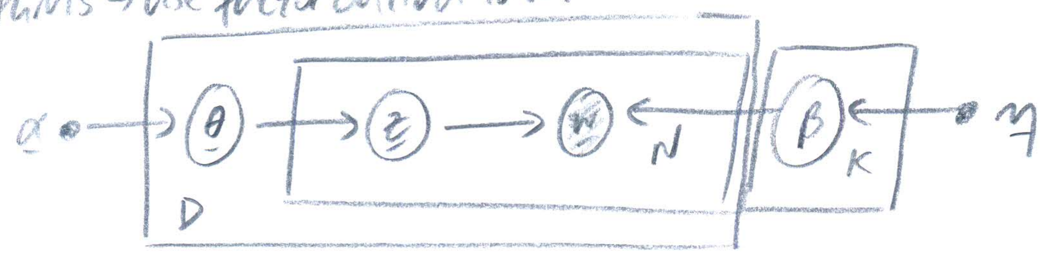
(*) Posterior inference results

- Familiar $\rightarrow$ recap   (A1) - review.

- (A) see allowed slides

(*) Joint likelihood   (A2) - check dim.

- PGMs $\rightarrow$ use factorization law!



$$p(\beta, \theta, z, w) = \prod_{k=1}^{K} p(\beta_k | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta)$$

(*) we want posterior of any of latent variables given observed words $\neq w$ and likelihood of the word.

(*) Inference and learning both intractable.

(A3) - Review the intractability to get a real sense of how powerful approx inference.

e.g. $p(\underline{\theta}_n|\underline{D})$ and $p(\underline{D})$

- No known technique for performing these techniques exactly.

(*) Approximate Inference

i) variational inference

- turns solution of an inference problem $\longrightarrow$ sol. of an optimization problem.

ii) MCMC

(*) Variational inference (cont)

(A4) - review logic

- consider generative model $p_\theta(\underline{x}|\underline{z})$, prior $p(\underline{z})$
- Joint distri $p_\theta(\underline{x},\underline{z}) = p_\theta(\underline{x}|\underline{z})p(\underline{z})$
- Assume variational distri $q_\phi(\underline{z}|\underline{x})$
- objective: maximise lower bound for log-likelihood:-

$$\log p(\underline{x}) = KL(q_\phi(\underline{z}|\underline{x}) \| p_\theta(\underline{z}|\underline{x})) + \int_{\underline{z}} q_\phi(\underline{z}|\underline{x}) \log \frac{p_\theta(\underline{x},\underline{z})}{q_\phi(\underline{z}|\underline{x})} d\underline{z} \quad ?$$

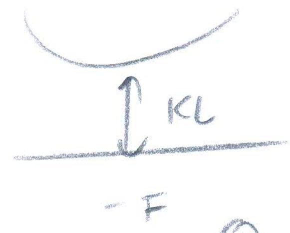
$$\geq \int_{\underline{z}} q_\phi(\underline{z}|\underline{x}) \log \frac{p_\theta(\underline{x},\underline{z})}{q_\phi(\underline{z}|\underline{x})}$$

$$:= \mathcal{L}(\theta;\phi;\underline{x})$$

Equivalently; minimise free-energy (upper bound on log-like.)
(surrogate on target)

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

(*) Distance between free energy and ...
(-F)



(*) $q_\phi(z|x) = p_\theta(z|x) \Rightarrow KL(q||p) = 0$

(*) <u>variational inference</u>

A5: <u>Review</u>

(*) Maximise variational lower bound:-

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

$$= \log p(x) - KL\left(q_\phi(z|x)\,||\,p_\theta(z|x)\right)$$

(W): A little abstract.

(*) <u>E-step</u>: maximise $\mathcal{L}$ wrt ϕ with θ fixed:-

$$\max_\phi \mathcal{L}(\theta;\phi;x)$$

If closed form sol. exist:-

do not set this to

$$q^*_\phi(z|x) \propto \exp\left[\log p_\theta(x,z)\right] \quad (\neq p(z|x) \text{ under V.I.})$$

(*) <u>M-step</u>: maximise $\mathcal{L}$ wrt θ; with ϕ fixed.

$$\max_\theta \mathcal{L}(\theta;\phi;x)$$

<u>Ex</u> use $q_\phi(z|x)$ i.e. an inference step on z (latent) given data x

- make sure q_ϕ is 'good' in sense of being 'close' to $p_\theta(z|x)$ using KL-divergence as a measure of <u>closeness</u>
- Why not just set $q_\theta(z|x) = p_\theta(z|x)$? $\rightarrow$ p_θ <u>intractable</u>

(*) <u>mean-field assumption (in topic models)</u>

- True posterior:- $p(\beta,\theta,z|w) = \dfrac{p(\beta,\theta,z,w)}{p(w)}$

(*) Break dependency using fully factorised distri:-

$$q(\beta, \theta, z) = \prod_{k} q(\beta_k) \prod_{d} q(\theta_d) \prod_{n} q(z_{dn})$$

- Product of non-coupled, individual distri (marginals)

Ex: Why does this make it things easier?

W:- Can optimise individually by breaking into subproblems

(A6): Review - not entirely sure of i) logic
ii) dimensionality

(*) Mean-Field Approx

(A7): Pick up during review. (Ex gone through quickly)

- Read/glance/skim papers

(*) Co-ordinate ascent algorithm for LDA

- Get an iterative program.
- (A8): Review pseudocode.

Ex: A lot of material to digest

- Next lectures ⟶ more examples of approx inference
- Give a 'grand theory' to unify/better understand.