10-708 L7 13/05/2020

L7 - Maximum likelihood learning of undirected GM

Ex: What is key algorithm for POGMs?

(A): EM

- 10-708: see lots of algorithms; develop taste and understanding
- POGMs: do inference on unobserved; then apply completely observed tools (heuristic)
- better researchers dig out foundations

E.g. EM as co-ordinate ascent algorithm (characterising it this way can place it in a class)

Ex: See rationale behind algorithm

Ex. use graphical models to pull together local structures

MLE for BNs:

(*) most important:- $\theta_{ijk}^{ML} = \dfrac{n_{ijk}}{\sum_{i,j',k} n_{ij'k}}$ (A: looks like counts/ empirical probability)

(however large GM is)

(*) Due to factorisability of DGM

Q: Does this apply to UGM?

MLE for undirected GMs

(*) UGM: Hammersley-Clifford means we can define a UGM in terms of a Gibbs distribution and partition function

$$p(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \qquad Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(x_n)$$

- Z - partition fn
  - normalisation constant of product of unnormalised potentials $\prod \psi_c(x_c)$

Ex: Suppose ψ (potentials) contains hidden parameters (e.g. strength of config per clique) and you want to estimate; given complete observations

Q Can you compute parameters within ψ easily?

Try to engage with line of thought

- A: No; you will have the following:- $\frac{1}{Z}\prod_{c\in C}\psi_c(\underline{x}_c, z)$

- So parameters will appear inside clique potential functions
- You will need to marginalise along the lines of

$$\prod_{c\in C}\sum_{z}\psi_c(\underline{x}_c, z)$$

(ψ_c not probability; but assume an analogous operation)

- So product-sum difficult for ML estimation

(*) Coupling is the answer

- $\psi_c(\underline{x}_c)$ ARE ✗
- Not complete the parameters wrong
→ you have written latent variables

EX: Nothing explicit to optimise against (if doing MLE); as you will have unknown parameters; and hence $\frac{1}{Z}\prod_{c\in C}\psi_c(\underline{x}_c)$ will be unknown

EX: Some graphical models structures can also be described by UGM.

(*) log-likelihood for UGMs with tabular clique potentials.

Sufficient stats:

- UGM (V, E): the no. times config $\underline{x}$ i.e. $\underline{X} = \underline{x}$ is observed in a data set $D = \{\underline{x}_1, \dots, \underline{x}_N\}$ can be represented as follows:-

Define

- $m(\underline{x}) = \sum_n \partial(\underline{x}, \underline{x}_n)$ (total count) OTT
- $m(\underline{x}_c) = \sum_{?} m(\underline{x})$ (clique count) OTT

W A

- Total counts - no. of time a configuration appears in dataset
- Clique counts - no. of times a particular configuration within a clique appears in the dataset

(*) clique counts obtained by marginalising our total counts (W)(A2)

Assume discreteness

(*)

log-likelihood:- $p(D|\theta) = \sum_c \sum_{\underline{x}_c} m(\underline{x}_c) \log \psi_c(\underline{x}_c) - N \log Z$

(W)(A3): check you understand how log-like is specified (quick)

ex: log-like:- Sum over all possible configurations of $\underline{x}$

- use delta function to clamp those values of $\underline{x}/\underline{x}_n$ (?) which are consistent with your observations of the data, count 1 everytime you see it.

- connects log-likelihood with ms (i.e. counts) obtained from data

(*) Do you understand how log-like and observations/sufficient statistics for UGMs are related?

(?)

(Q): What is θ (parameters?) (so you remember L3!)

(W): (A4): Refresh param. of UGMs!

- we are computer scientists, not mathematicians
- Do not matter to theoreticians (actually dealing with messiness)
- You were close → CPDs; but do not obey constraints of prob.
- An unnormalised table of nos $\psi_c(\underline{x}_c)$ - $\underline{x}_c$ associated with a no. e.g α

(*) Derivative of LL:-

- standard calculus:

1st term $\dfrac{\partial l_1}{\partial \psi_c(\underline{x}_c)} = \dfrac{m(\underline{x}_c)}{\psi_c(\underline{x}_c)}$

2nd term: (W)(A5) - Review this (quick)

---

conditions on clique marginals

- get optimal ψ_c^* → find it vanishes

(*) At ML setting of parameters; for each clique, model marginals equal to observed marginals (empirical counts)

EX: Only get marginal probability of a clique $\hat{p}_{MLE}(\tilde{x}_c)$; we want the potential function (estimates of) of each clique.

- These are not the same in UGMS (✓)

(*) Only provides condition that must be satisfied when we have ML param; does not specify how to get ML param.

EX: serious work into doing this

MLE for UGMS

EX: Previous iterations relied on these concepts (decision tree style questions)

- triangulated
- clique potentials defined on maximal cliques
- full tables or compact

} - skip - see Koller for historical account

---

2 workhorse algorithms (most insightful)

- IPF (Iterative proportional fitting) - MRFs tabular pot. — nos. behind every config
- GIS (Generalised iterative scaling) - MRFs with features potential
- EX: Key is how the differences in problem scopes yield ~~to~~ differences in algorithmic approach
  Algebraic tricks → make problem easier (significantly)

IPF

- Identity from LL optimisation → anti-climactic
- How to recover from this?

(*) From LL: $\dfrac{\partial l}{\partial \psi_c(x_c)} = \dfrac{m(x_c)}{\psi_c(x_c)} - N \dfrac{p(x_c)}{\psi_c(x_c)}$

(W)(A6): - Review this algo.

$$p^*_{MLE}(x_c) = \frac{m(x_c)}{N} = \hat{p}(x_c)$$

(*) Derive:- $\dfrac{\tilde{p}(x_c)}{\psi_c(x_c)} = \dfrac{p(x_c)}{\psi_c(x_c)}$

· Turn identity into a fixed point equation (now identity with time component) for ψ_c

$$\psi_c^{(t+1)} = \psi_c^{(t)}(x_c)\,\frac{\hat{p}(x_c)}{p^{(t)}(x_c)}$$

(*) update fn: $\dfrac{\hat{p}(x_c)}{p^{(t)}(x_c)}$ - proportion of empirical marginal (countable from data) over current version of estimated marginal; based on your model. (derivable from $\psi_c^{(t)}(x_c)$)

(*) In UGM; even with observed data, have to do inference

Additional question: i) Does it converge etc.?

Properties of IPF updates:

IPF is a fixed point program over time; but also over potential functions

(*) n(?): over potentials

(*) A co-ordinate ascent algorithm; attaining an optima in a part. direction when other directions fixed.

- convergence somewhere.
- (*) Also known as I-projection (distn from one space to another where only one potential is allowed to change)

- over space of possible distn families (via attained via max-entropy)

(*) understood
via KL divergence view → comes up in V.I/D.L (Jordan 11)

KL divergence view

- ML can be reframed as KL divergence
- coordinate ascent charac. of IPF through KL divergence (via info theory)

$$\max \ell \Leftrightarrow \min KL(\hat{p}(\underline{x}) \| p(\underline{x}|\underline{\theta})) = \sum_{\underline{x}} \hat{p}(\underline{x}) \log \frac{\hat{p}(x)}{p(\underline{x}|\underline{\theta})}$$

- Partition arguments of distri into:-

X_c and $X_{c\backslash}$ — $c\backslash$ - complement of c

- To carve out a parti. potential clique

(*) combine KL with conditional chain rule (L)(95): review ✓

IPF minimises KL divergence

(*) changing ψ_c (clique potential) has no effect on c.d.
(2nd term unaffected)

(*) $$KL(\hat{p}(\underline{x}) \| p(\underline{x}|\underline{\theta})) = KL(\hat{p}(\underline{x}_c) \| p(\underline{x}_c|\underline{\theta})) + \sum_c \hat{p}(\underline{x}_c) KL(\hat{p}(\underline{x}_{c\backslash}|\underline{x}_c) \| p(\underline{x}_{c\backslash}|\underline{x}_c))$$

- i.e. setting $p(\underline{x}_c) = \hat{p}(\underline{x}_c)$

(L)(96): quick review

- IPF
- start with random guess of potential nos. $\psi_c^{(0)}(\underline{x}_c)$
- multiply by a ratio $\dfrac{\hat{p}(\underline{x}_c)}{p^{(t)}(\underline{x}_c)}$ (proportional)

- convexity only qualifies whether local or global (in this context)

- initialise random no. generator 100 times and run (to deal with converg/local/global)

-