10-708 06/05/2020

L5 - Paramete estimation (s.2020)

- EX: Paramete estimation in PGMS (a clean reference point)

- Start with named/indexed r.v.s.
- And training data (IID); fully observed - every r.v. has instantiation

1.) Structure learning
   - in principle, possible to learn structure from data
   - often experts.

2) Paramete estimation
   UGM - NOS in CPT or potential function values
   DGM

EX: learning completely observed UGM - fairly trivial

POGMs (partially obs.)
- directed - focus on this

Estimation principles

MLE → classical setting statistics

EX: ML has 'gone beyond this'; present a universal, standardised view that unifies these.

e.g. reinforcement learning - intrinsic/extrinsic
     adversarial learning - adversarial score

- Traditionally, characterise learning/param est in terms of statistical consistency etc.
- in more modern engineering ML → may be compromised

- Simplest case
- COGMS where structure is known
- Parameter learning for BN
-

(*) Analytically write down loss fn → likelihood of data as a function of the parameters

- Probability (likelihood)? as product of many local terms

EX: Point you to further reading

- Building blocks of GM:

- Single node GM (e.g. root node in tree) → supp slides W A1

- instances of exponential family distri

(*) ultimately; paramete (probability) is empirical frequency count

---

2 node graphical models → W A2

↳ supp. slides

EX: Patterns underlying paramete learning for other use

Exponential family

EX: class interested in building blocks.

$$p(x|\eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}$$

$T(\cdot)$ and $\eta(\cdot)$ most important terms; note dot product

$A(\cdot)$ - log normaliser

EX: Estimation of (canonical) parameters η only require $T(\cdot)$ (i.e. sufficient statistic)

· Examples - MVG

- canonical, moments related

- μ and Σ - moment parametes

W A1 → Review exp

Exponential family representation

vec$(\cdot)$ - ~~linearise~~ / turn the entity into a vector

η and $T(\cdot)$ - of same dimensionality

Multinomial example

- constrained parameters in Multinomial (degrees of freedom) [sum to unity]
- Hence (K-1) in summation

Exponential family rep.

Q: Why do we go for exponential family reps?

① - Data and parameters clearly grouped into 2 terms — through η / through $T(\cdot)$

② Moment generating function.

$$\frac{dA}{d\eta} = \mathbb{E}[T(x)] \qquad \frac{d^2A}{d\eta^2} = Var[T(x)]$$

- Gives standard operator that yield n^{th} order moments (from derivatives of log-normaliser)

EX: Moments important → characterise

③: Relationship between moment and natural parameters

(*) Key - Review moment and canonical param relation.

MLE for exponential family

- only differences between distri in exp. family are form of η and $T(\cdot)$ i.e. canonical param and suffic stat.
- IID data → log-likelihood → optimise (set 1^{st} order moment to 0)
- moment matching

EX: Exponential family exposes the relationship between data (through $T(\cdot)$) and parameters ~~through~~ (through η) (in a linearly dependent fashion)

- Gives info about transformations of data or forms of data we need to worry about to preserve uniqueness and identity of distri.
- E.g. only store sufficient statistic of data

- 3 ways of conceptualising relation between: (dependencies)

X (data) $\quad T(x)$ (suff. statistic) $\quad \theta$ (param)

- Bayesian: - Draw conclusions on parameters given data
  - dependency of parameters from data
  - use posterior $p(\theta | T(x), x)$

Frequentist - Data generated from unknown true value of param.
  - Parameters impact data only through suff. stat $p(x | T(x), \theta)$

- ~~At~~ influence flows through $T(\cdot)$ for both Bayesian, frequentist (due to exponential family)

· Neyman factorisation theorem: W A3 → WIKI - check you understand eq.

- exposes sufficiency of $T(x)$ for parameter θ.
- $T(x)$ d-separates X and θ

(*) Density estimation for single r.v. for many diff distr family
  - use sufficient statistic, moment matching, exp. family

· Move onto 2 nodes

- Generalised instance → GLIM

EX: Builds on knowledge of exponential family

- Discrim - logistic regression ; SVMs

  LDA → No it's generative

- logistic regression - $p(y=1|x) = \dfrac{1}{1+e^{-\theta^T x}}$

- These are GLIMS
  - But sigmoid → non linearity (taken care of by blanket function)
  - contains linear rel. ; so we use linear techniques with above.

Commonality (*):- $\mathbb{E}_p(Y) = \mu = f(\theta^T x)$

- $p(Y|f(X))$ → conditional density of X → use exponential family distn

- $f(\cdot)$ is a response function → ~~tt~~ X treated in a linear way (dot product)

(W) (A4): check you understand formulation of GLIM (*)

(*) Different choices of p() and f() - cover all of 2 node GMs.

Linear Regression

Logistic

MRFs : No y; but exponential family distn

RBMs

CRFs

} (W) (A5): check that you understand these as GLIMs

GLIM (cont.)

- simple modelling principle: (for many instant.)

(*) Begin from data $\underline{x}$
Assume set of parameters $\underline{\theta}$ (to be estimated) } interact linearly to form signal 集约.

(*) Signal ξ turned into mean parameter of cond distn of output via response function f()

(*) Mean parameters and canonical parameters related by ψ (inverse transform for relating distribution instance params and canonical funct. form)

(*) Use exponential family to get y.

- mechanical pipeline

(*) Work in exponential form → lots of key results

- Ex: clear relationship between $f(\cdot)$ and $\psi(\cdot)$ - allows 'cancellation'

$f = \psi^{-1}(\cdot)$

- Analytical simplif (?) w (A6)

MLE for GLIMS → natural response

- $f(\cdot)$ and $\psi(\cdot)$ "cancel" → allows simplistic def. of cond. likelihood of output given input.

W(17): check this reasoning:- $f(\cdot)$ and $\psi^{-1}(\cdot)$

Yields online learning for canonical GLIM and stochastic gradient ascent

(Q): How is $\psi(\cdot)$ chosen?

- Given a distn instance, use default (?)
- can specify any (?)
- some don't choose $\psi(\cdot)$

(*) S.G.A can be used for any GLIM model (slow)

- Batch learning for GLIM
- Best is Newton's method → requires knowledge of ψ function
- Requires computation of Hessian
- We already have a library of canonical response functions

IRLS / Newton-Raphson

- Requires Hessian (inverse) and gradient of loss.
- Reframe update rule to see how it relates to LS loss.

Ex: Spirit is exponential family and GLIMS use universally (one node and 2 node building blocks for GM).

W(18): check you understand full GLIM / exp formulation of logistic, linear.

MLE for general BNS:-

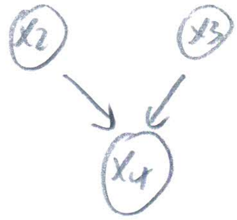
(*) Assume global indep. of param; nodes fully observed, decompose BN by decomposing log-likelihood into a sum of local terms, one per node.

Decomposable likelihood of BN

- graphical illustration

- illustrates analytic decomposition, use of GLM/exp. methods

Ex: How about 2 parents



- GLM/exp?
- Additive $X_1 + X_2$
- Multip. $X_1 X_2$
- multiplexer - (previously pop.)

- Multiplexer function
  - $\prod_{i=3}^{k} X_i^{\delta(X_2, i)}$

- combine inputs asymmetrically

$$\mathbb{I}\left(\prod_{i=3}^{k} X_i^{\delta(X_2, i)}\right) \longrightarrow \text{still GLM}$$

MLE for BNs with tabular CPDs

- Get simple estimators (via same procedure)

S.2020 lecture 5 is a hybrid of S.2019 L5 and L6 material

- so this is technically S.2019 L6 (POGM param est.)

- POGMs (partially observed GMs)

Ex: too useful in practical situations e.g. speech HMMs (latent/unobs words)

biolog ev. clustering (latent clusters)

Mixture models

- observe $\underline{x}$ as 2d co-ord., no label (z)
- estimate $p(z|\underline{x})$ generatively via $p(\underline{x})\, p(\underline{x}|z)$

Decided to pause here: concluded run of L6 S.2019