Ex: 4th year PHD did not know how to run K-means
- You want to know theorem; but you have to get it working (OT)
- (w): implement these(?)

- Potential functions $\psi_c(\mathbf{x}_c)$
- Some issues: - spell-checking example
  - Build affinity models of charact. streams
  - e.g. consec appearances of 3 streams
  - use this criterion to score likelihood

$\psi_c(\mathbf{x}_c) = \psi_c(x_1, x_2, x_3)$

'zyz' and 'ink'

- Define potential function over a triplet of characters
- 26^3 different features and put a function on it?
- 26^5

(*) Above shows infeasibility of tabular potentials (akin to enumerat. of joint)

Feature based clique pot.

(w): note instantiation of cliques in practice

(*) Features

(?): Find a way of appropriately compressing the granularities

- Handcrafted feature design → role of human knowledge in A.I
- use feature-engin to save on rep. cost?
- Distinct from current, overcomplete ML models (e.g. gigantic placeholder)
- Ex: Against the idea of human knowledge being ignored in ML

(*) microportentials fk OR
- Have k features and weights defined over 3 charact. potential

- $\psi_c(c_1, c_2, c_3) = \exp\left\{\sum_{k=1}^{K} \theta_k f_k(c_1, c_2, c_3)\right\}$

(*) overall potential (clique) is exp weighted sum of micropot.

(*) micropotent. distinct from tabular potentials.

(*) K parameters over K features → more compact

Combining features

(*) sliding window/overlapping sliding window

(*) Note how we can modify standard of Gibbs. rep. for exp.

- Allows use of exponential /GLIMS

- (?) : Not entirely clear how to apply IPF in this case due to coupling of estimated θ_k and designed $f_k(c_1, c_2, c_3)$

MLE of feature based UGMs:

- Scaled likelihood:- $\hat{l}(\theta; D) = l(\theta; D/N) = \frac{1}{N} \sum_n \log p(x_n | \theta)$

$$= \sum_x \hat{p}(x) \log p(x | \theta)$$

$$= \underbrace{\sum_x \hat{p}(x) \sum_i \theta_i f_i(x)}_{(i)} - \underbrace{\log Z(\theta)}_{(ii)}$$

(*) Calculus + derivatives → not fruitful.

Ex: Nonlinearities cause issues (e.g. log/inversion)

- Linearise it so argument can be exposed to linear attack

- log has linear upper bound

- $\log Z(\theta) \le \mu Z(\theta) - \log \mu - 1$

- Bound holds $\forall \mu$: $\mu = Z^{-1}(\theta^{(t)})$

fixed point it. strategy

assume this [there is a previous version of Z]

(*) GIS derivation (1)(17): Review

- Define $\Delta\theta_i^{(t)} = \theta_i - \theta_i^{(t)}$ and introduce

- still nasty: → every

(*) Note exp of weighted sum: - $\exp\{\sum_i \Delta\theta_i^{(t)} f_i(\underline{x})\}$

EX: We make distinction between weight θ_k and $(\Delta\theta_i^{(t)})$ and features $f_i(\underline{x})$; algebraically the same.

- Treat f_i as weights; $\Delta\theta_i^{(t)}$ as arguments

- (*) impose prob. constraints (normally applying to weights) to f_i our assumed "weights".

relaxation

(*) exp(·) is convex → use Jensen's

- Algebraic trick often used in ML

- getting $\sum_i f_i(\underline{x}) \exp(\Delta\theta_i^{(t)})$ → only linearly coupled with others.

(*) use low bound of scaled LL: - GIS

(*) use calculus: -

(*) giving update steps: -

(*) Note: - $\dfrac{\sum_{\underline{x}} \hat{p}(\underline{x}) f_i(\underline{x})}{\sum_{\underline{x}} p^{(t)}(\underline{x}) f_i(\underline{x})}$

"weighted sum of features by emp. prob. $\tilde{p}(\underline{x})$"

"features weighted by inferred probability $p^t(\underline{x})$"

(*) Iterative re-scaling → connection with IPF.

(*) Summary GIS/IPF

- fixed point iterations on LL obj.

- one for tabular, feature based

-

(*) Where does exponential come from? (Ⓠ: A move from Gauss → Gibbs?)

Ⓛ A7: Review exp. family form (LS)

(*) Note at MLE; expectations of sufficient statistics under model match exp. feature average'

- Note eq.

(*) Begin with exp. family → get a consequence.

- Reverse rationale
- Maximum entropy

- impose constraint on distri: so you can't give me arbitrary ones expectat. of feature must match sample av.
- Fixed feat. exp: $\sum_x p(x) f_i(x) = \alpha_i$ (from data)

- encode few assumpt. about model as poss.
- entropy as amt of randomness / amt of assumptions made

$$\max_p H(p(\underline{x})) = -\sum_{\underline{x}} p(\underline{x}) \log p(\underline{x})$$

$$\text{s.t.} \quad \sum_{\underline{x}} p(\underline{x}) f_i(\underline{x}) = \alpha_i$$

$$\sum_{\underline{x}} p(\underline{x}) = 1 \longrightarrow p(\underline{x}|\underline{\theta}) = \frac{1}{Z(\underline{\theta})} \exp\left\{\sum_i \theta_i f_i(\underline{x})\right\}$$

(*) Variational definition: - define a distrib. as a solution to a constrained optimisation problem.

Ⓛ A8: Review Lagrangian sol.

(*) Natural consequence gives: an exponential family distri.

(*) Benefit of information theoretic principles in ML → Ⓛ: explore

- More general max entropy method
- incorporate prior distri on $\underline{x}$; reference it. $(u(\underline{x}))$
- Estimated distri has least addit. assumptions from priors
- use KL-divergence rather than entropy

$$\min_{p} \; KL(p(\underline{x}) \| h(\underline{x})) = \sum_{\underline{x}} p(\underline{x}) \log \frac{p(\underline{x})}{h(\underline{x})} = -H(p) - \sum_{\underline{x}} p(\underline{x}) \log h(\underline{x})$$

$$\text{s.t.} \; \sum_{\underline{x}} p(\underline{x}) f_i(\underline{x}) = \alpha_i$$

$$\sum_{\underline{x}} p(\underline{x}) = 1$$

$$\Longrightarrow p(\underline{x}|\underline{\theta}) = \frac{1}{Z(\underline{\theta})} h(\underline{x}) \exp\left\{\sum_i \theta_i f_i(\underline{x})\right\}$$

(*) Constraints from data

Q. Where do constraints α_i come from?

- data itself is the constraint
- (h) (PG): Automatic consistency?

- Geometric intep; general process:-

Either:-

1) Assume all exponential family distns as "model":-

$$\mathcal{E} = \left\{p(\underline{x}) : p(\underline{x}|\underline{\theta}) = \frac{1}{Z(\underline{\theta})} h(\underline{x}) \exp\left\{\sum_i \theta_i f_i(\underline{x})\right\}\right\}$$

OR

2) Assume all distribs satisfying moment constraints

$$\mathcal{M} = \left\{p(\underline{x}) : \sum_{\underline{x}} p(\underline{x}) f_i(\underline{x}) = \sum_{\underline{x}} \hat{p}(\underline{x}) f_i(\underline{x})\right\}$$

do not acknowledge roots

- Information geometry Pythagorean theorem:-

(*) Summary

→ inspires V.I; deep gen. models

(*) Exp family viewed as a sol. to variational exp → maximum entropy

(*)

/ 2020 (

Supplementary → structure learning (see supp.)

Case study: CRFs (Lafferty) - at CMU
- insight of experienced modelling

- Lafferty paper → impressive ; clear rationale, motivation

(*) 'local normalisability is a double-edged sword
- makes computing simple'

(?) an - in context of HMM

(*) What you want is global normalisability

- use scores rather than enforcing local normalisability
- use potentials

$$\exp\{\sum \theta_i f_i\}$$

- features corresp. to nodes
- use human knowledge for features

- 'Art of modelling' → totally comprehens.