10-708 20/05/2020

L5 - Parameter estimation for PGMs - review

(*) Review slides → note key eq.
(*) use readings to supplement areas of uncertainty

(*) General parameter est. setup
- Assume structure of graph G is fixed / given in advance
- estimate parameters from IID dataset $D = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\}$
- N training instances
- Each training instance $\underline{x}_n = \begin{pmatrix} x_{n,1} \\ \vdots \\ x_{n,M} \end{pmatrix}$ $\underline{x}_n \in \mathbb{R}^M$ - Each dim. of $\underline{x}_n$ corresponds to a realisation of an r.v i.e. a node

(*) completely observable
- $x_{n,i}$ is known $\forall$ $n = 1, \dots, N$ and $i \in 1, \dots, M$

(*) partially observable

$\exists i : x_{n,i}$ is not observed.

- log-likelihood (function of param):- (general BN) π_i - parents of node i

$$l(\underline{\theta}; D) = \log p(D|\underline{\theta}) = \log\left(\prod_{n=1}^{N}\left(\prod_{i=1}^{M} p(x_{n,i} | \underline{x}_{n,\pi_i}, \theta_i)\right)\right)$$

$$= \sum_{i=1}^{M}\sum_{i=1}^{N} \log p(\underline{x}_{n,i} | \underline{x}_{n,\pi_i}, \theta_i)$$

(*) Exponential family distr. (Murphy 2012)
- For a ~~numeric~~ vector r.v. X :- (~~scalar~~?) (*)

$$p(x|\eta) = h(\underline{x}) \exp\{\eta^T T(\underline{x}) - A(\eta)\} = \frac{1}{Z(\eta)} h(\underline{x}) \exp\{\eta^T T(\underline{x})\}$$

- is the exponential family distr with eq :-

(*) canonical param η
(*) sufficient statistic $T(\underline{x})$
(*) log normaliser $A(\eta) = \log Z(\eta)$

(*) exp family + GLMs → makes many models instances of this general form

Examples:-

i) MVG:

- $\underline{x} \in \mathbb{R}^K$

$$p(\underline{x}|\underline{\mu},\underline{\Sigma}) = \frac{1}{(2\pi)^{K/2}|\underline{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})^T\underline{\Sigma}^{-1}(\underline{x}-\underline{\mu})\right\} \quad ^{(i)}$$

$$= \frac{1}{(2\pi)^{K/2}|\underline{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}tr\left(\underline{\Sigma}^{-1}\underline{x}\underline{x}^T\right) + \underline{\mu}^T\underline{\Sigma}^{-1}\underline{x} - \frac{1}{2}\underline{\mu}^T\underline{\Sigma}^{-1}\underline{\mu} - \log|\underline{\Sigma}|\right\}$$

(i) obtained by expanding; applying the trace trick to quadratic form $\underline{x}^T\underline{\Sigma}^{-1}\underline{x}$

- As $\underline{x}^T\underline{\Sigma}^{-1}\underline{x} = tr(\underline{x}^T\underline{\Sigma}^{-1}\underline{x}) = tr(\underline{x}\underline{x}^T\underline{\Sigma}^{-1}) = tr(\underline{\Sigma}^{-1}\underline{x}\underline{x}^T)$

Exponential family

$$vec\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = \begin{bmatrix} a \\ c \\ b \\ d \end{bmatrix}$$

$$\eta = \begin{bmatrix} \underline{\Sigma}^{-1}\underline{\mu} \\ -\frac{1}{2}vec(\underline{\Sigma}^{-1}) \end{bmatrix} = \begin{bmatrix} \eta_1 \\ vec(\eta_2) \end{bmatrix} ; \quad \eta_1 = \underline{\Sigma}^{-1}\underline{\mu}, \quad \eta_2 = -\frac{1}{2}\underline{\Sigma}^{-1}$$

$$T(\underline{x}) = \begin{bmatrix} \underline{x} \\ vec(\underline{x}\underline{x}^T) \end{bmatrix}$$

$$A(\eta) = \frac{1}{2}\underline{\mu}^T\underline{\Sigma}^{-1}\underline{\mu} + \log|\underline{\Sigma}| = -\frac{1}{2}tr(\eta_2\eta_1\eta_1^T) - \frac{1}{2}\log(-2\eta_2)$$

$$h(\underline{x}) = (2\pi)^{-K/2}$$

- $K+K^2$ parameters; but lower due to symmetric PSD $\underline{\Sigma}$

ii) Multinomial

$$\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_M \end{pmatrix} \quad \underline{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_M \end{pmatrix}$$

- $\underline{X} \in$ Multinomial$(\underline{X}|\underline{\pi})$

- Bishop(?) (EX is incorrect) (not sure what lecture notes are refering to)

Jordan (2003): X is a collection of integer-valued r.v.s. representing event counts where X_k is no. of times the k^{th} event occurs in n independent trials.

(*) moment generating properties of exp family

(scalar form.)

· The moments of the relevant distr → obtained via derivatives of log normalisation function $A(\eta) = \cancel{\log(1+e^{\eta})} = \log z(\eta)$

(?) confused by dimensionality of (*)

(*) $$\frac{dA}{d\eta} = \frac{d}{d\eta}\log z(\eta) = \frac{1}{z(\eta)}\frac{d}{d\eta}z(\eta)$$

$$= \frac{1}{z(\eta)}\frac{d}{d\eta}\int h(x)\exp\{\eta T(x)\}\,dx$$

(I) ↓

$$= \int T(x)\,\frac{h(x)\exp(\eta T(x))}{z(\eta)}\,dx$$

(*)

$$= \mathbb{E}[T(x)]$$ (with respect to?)

(*) $$\frac{d^2A}{d\eta^2} = \int T^2(x)\frac{h(x)\exp\{\eta^T T(x)\}}{z(\eta)}dx - \int T(x)\frac{h(x)\exp\{\eta^T T(x)\}}{z(\eta)}dx\,\frac{1}{z(\eta)}\frac{d}{d\eta}z(\eta)$$

$$= \mathbb{E}[T^2(x)] - \mathbb{E}^2[T(x)]$$

$$= \mathrm{Var}(T(x))$$

0/

(*) Take derivatives of log-normaliser

- q^{th} derivative → q^{th} centred moment

$\frac{dA(\eta)}{d\eta}$ - mean , $\frac{d^2A(\eta)}{d\eta^2}$ - variance

(*) sufficient statistic - vector ⟹ partial deriv.

(I) clarity on derivation

$$\frac{dA}{d\eta} = \frac{d}{d\eta}\left\{\log\int\exp\{\eta T(x)\}h(x)\,dx\right\} = \frac{\int\frac{d}{d\eta}\exp\{\eta T(x)\}h(x)\,dx}{\int\exp\{\eta T(x)\}h(x)\,dx}$$

(*) ~~If p~~ If π_k the probability of k^{th} event occuring on any given trial:-

$$p(\underline{x}|\underline{\pi}) = \frac{n!}{x_1! x_2! \dots x_m!} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_m^{x_m}$$ (pmf)

$$\Rightarrow p(\underline{x}|\underline{\pi}) = \exp\left\{\sum_{i=1}^{m} x_i \ln \pi_i\right\}$$ (trick)

(*) Issue:- factor in probability constraint $\sum_{i=1}^{m} \pi_i = 1$

↳ $A(\eta) = 0$?

(*) Parametrise with only $(m-1)$ components of $\underline{\pi}$:-

$$\begin{aligned} p(\underline{x}|\underline{\pi}\setminus\pi_m) &= \exp\left\{\sum_{i=1}^{m} x_i \ln \pi_i\right\} \\ &= \exp\left\{\sum_{i=1}^{m-1} x_i \ln \pi_i + \left(1 - \sum_{i=1}^{m-1} x_i\right) \ln\left(1 - \sum_{i=1}^{m-1} \pi_i\right)\right\} \\ &= \exp\left\{\sum_{i=1}^{m-1} \ln\left(\frac{\pi_i}{1 - \sum_{i=1}^{m-1} \pi_i}\right) x_i + \ln\left(1 - \sum_{i=1}^{m-1} \pi_i\right)\right\} \end{aligned}$$

Exp family rep

$$\eta = \begin{bmatrix} \ln\left(\frac{\pi_i}{\pi_m}\right) \\ \\ 0 \end{bmatrix}$$

i.e. η has $(m-1)$ components $\eta_i = \ln\left(\frac{\pi_i}{\pi_m}\right)$ and last component $\eta_m = 0$

$$T(\underline{x}) = \underline{x}$$

$$A(\eta) = -\ln\left(1 - \sum_{i=1}^{m-1} \pi_i\right) = \ln\left(\sum_{i=1}^{m} e^{\eta_i}\right)$$

$$h(\underline{x}) = 1$$

(*) Note $\pi_i = \dfrac{e^{\eta_i}}{\sum_{j=1}^{m} e^{\eta_j}}$ (softmax) $\pi_i = \text{softmax}_i(\eta)$

$$= \frac{\int T(x)\exp\{\eta T(x)\} h(x)\, dx}{\int \exp\{\eta T(x)\} h(x)\, dx} = \int T(x)\exp\{\eta^T T(x) - A(\eta)\} h(x)\, dx$$

?

O/S 1

$$= \mathbb{E}[T(X)]$$

(*) Moment and canonical parameters.

(*) Exponential families can have {canonical parametrisation (via η)
moment parametrisation (via μ)

- $\frac{dA(\eta)}{d\eta} = \mathbb{E}[T(x)] = \mu$; $\frac{d^2A(\eta)}{d\eta^2} = \mathrm{Var}(T(x)) > 0$ (variance properties → non-neg)

(*) $A(\eta)$ is convex function

(*) convexity ⇒ one-to-one rel between argument η and first derivative.

(*) yields an invertible mapping :- (*)

$$\eta = \psi(\mu)$$

(*) Moment matching, MLE for exponentials

O/S 2 - dimensional clarity

- IID data:

- Log likelihood

$$l(\eta; D) = \log \prod_{n=1}^{N} h(x_n)\exp\{\eta^T T(x_n) - A(\eta)\}$$

$$= \sum_{n=1}^{N} \log h(x_n) + \eta^T\left(\sum_{n=1}^{N} T(x_n)\right) - NA(\eta)$$

$$\nabla_\eta l = \sum_{n=1}^{N} T(x_n) - N\nabla_\eta A(\eta) = 0$$

$$\Rightarrow \nabla_\eta A(\hat{\eta}) = \frac{1}{N}\sum_{n=1}^{N} T(x_n)$$

- define $\mu = \mathbb{E}[T(\underline{x})]$

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} T(\underline{x}_n)$$

(*) moment matching $\Rightarrow$ can infer canonical param via $\hat{\eta}_{ML} = \psi(\hat{\mu}_{ML})$ and vice versa

---

(*) Sufficiency + Neyman factoris.

Bayesian: $X \longrightarrow T(x) \longrightarrow \theta$ $\quad p(\theta \mid T(x), x) = p(\theta \mid T(x))$ (prob. spec.)

i.e. $\theta \perp\!\!\!\perp X \mid T(x)$ (CI rel. spec.)

Freq: $X \longleftarrow T(x) \longleftarrow \theta$ $\quad p(x \mid T(x), \theta) = p(x \mid T(x))$

~~i.e.~~ i.e. $(\theta \perp\!\!\!\perp X \mid T(x))$

O/S 3

- Reconcile/review in context of 36-705

Neyman fact:

$X - T(x) - \theta$

- via UGM formalism: - $T(x)$ is sufficient for θ

$$p(x, T(x), \theta) = \psi_1(T(x), \theta)\psi_2(x, T(x))$$

- proport/norm constant; absorbed into one of ψ_1/ψ_2 potentials

$$\Rightarrow p(x \mid \theta) = g(T(x), \theta)\, h(x, T(x))$$

- $T(x)$ - determ function of x; drop from LHS; divide by $p(\theta)$

- for given g and h (functions).

Remarks

(*) can specify either C.I statements or prob. versions

(*) Bayesian intuitions

- θ is an r.v. $\longrightarrow$ can make C.I. statements involving θ

(*) frequentist intuitions

- Treat θ as a label rather than r.v.; $T(X)$ is sufficient for θ if the conditional distrib of X given $T(X)$ is not a function of θ.

(*) Both imply a factorisation of $p(x|\theta)$

(*) Neyman factorisation $\longrightarrow$ frequentist definition of sufficiency

(*) Sufficiency in this context means $T(x)$ is sufficient for θ if:-

$$\theta \perp\!\!\!\perp X \mid T(X)$$

Jordan (2003)

- 8.1.8. ML and KL divergence
- A general rel. between ML and KL divergence (not spec. to exp.)
- Necessary for later lec. material L6, L7
- Statistical interp of KL divergence to illus. rel. between KL and exp. family

(*) Empirical distn: $\hat{p}(\underline{x})$

- Places a point mass at each data point $\underline{x}_n$ in D (dataset)

(discrete)

(*) Empirical distn: $\hat{p}(\underline{x}) := \frac{1}{N}\sum_{n=1}^{N} \partial(\underline{x}, \underline{x}_n) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(\underline{x} = \underline{x}_n)$

∂ - Kronecker delta in cont. case.

(*) Sum/integrate $\hat{p}(\underline{x})$ against a function of $\underline{x}$, i.e. we evaluate for at each point $\underline{x}_n$

(*) scaled log likelihood: (also, cross entropy of $\hat{p}(\underline{x})$ and $p(\underline{x}|\underline{\theta})$)

$$\sum_{\underline{x}} \hat{p}(\underline{x}) \log p(\underline{x}|\underline{\theta}) = \sum_{\underline{x}} \left(\frac{1}{N}\sum_{n=1}^{N} \partial(\underline{x}, \underline{x}_n) \log p(\underline{x}_n|\underline{\theta}) \right)$$

$$= \frac{1}{N}\sum_{n=1}^{N} \sum_{\underline{x}} \partial(\underline{x}, \underline{x}_n) \log p(\underline{x}_n|\underline{\theta})$$

detrickifying

$$= \frac{1}{N}\sum_{n=1}^{N} \log p(\underline{x}_n|\underline{\theta})$$

$$= \frac{1}{N} l(\underline{\theta}|D)$$

(*) Note; the scaled log likelihood (by factor $\frac{1}{N}$) is equivalent to the
(*) cross-entropy between the empirical distri $(\hat{p}(\underline{x}))$ and the model to $p(\underline{x}|\underline{\theta})$

- See result for continuous

- KL divergence between empirical; model :-

$$D(\hat{p}(\underline{x}) \| p(\underline{x}|\underline{\theta})) = \sum_{\underline{x}} \hat{p}(\underline{x}) \log \frac{\hat{p}(\underline{x})}{p(\underline{x}|\underline{\theta})}$$

$$= \sum_{\underline{x}} \hat{p}(\underline{x}) \log \hat{p}(\underline{x}) - \sum_{\underline{x}} \hat{p}(\underline{x}) \log p(\underline{x}|\underline{\theta})$$

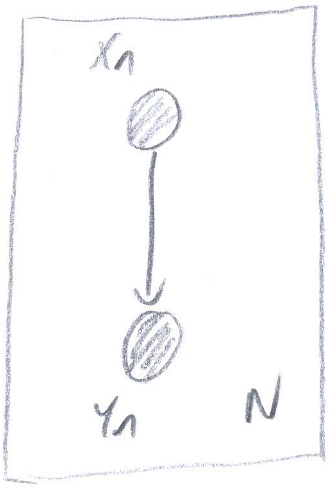
$$= \underbrace{\sum_{\underline{x}} \hat{p}(\underline{x}) \log \hat{p}(\underline{x})}_{(I)} - \underbrace{\frac{1}{N} \ell(\underline{\theta}|D)}_{(II)}$$

(I) - independent of θ
- value of $\underline{\theta}$ that minimises LHS is the value of $\underline{\theta}$ that maximises the RHS

(*): minimising KL divergence between the empirical distri and model distri is equivalent to maximising the likelihood

---

(*) Generalised Linear Models (GLM) - linear regression/classification



- covers linear reg. / discriminative linear classification.

(*) Both LR / LC → both assume a rep. for conditional expectation of Y.

(*) $\mu = f(\underline{\theta}^T \underline{x}) = \mathbb{E}_{Y \sim p(Y|f(x))}[Y]$

(*) LR: $f(\cdot)$ - identity
LC: $f(\cdot)$ - sigmoid (logistic)

(*) Also: endow Y with a particular cond. prob. distribution, with μ as a parameter.

(W): Remember JP $\longrightarrow$ Columbia X ML (prob. interp of ML for LR!)

(*) LR - Gaussian LC - Bernoulli / Multinomial.

(*) Generalised linear model framework

- 3 assumptions on $p(y|\underline{x})$:-

1. observed input $\underline{x}$ enters into model via linear comb. $\xi = \underline{\theta}^T \underline{x}$
2. conditional mean μ rep as a function $f(\xi)$ of the linear combination ξ where f is known as the response function
3. observed output y is assumed to be characterised by exp. family with conditional mean μ.

(R)

$$\begin{tikzcd} \underline{\theta} \arrow[dr] & & & \\ & \xi \arrow[r, "f"] & \mu \arrow[r, "\psi"] & \eta \\ \underline{x} \arrow[ur] & & & \end{tikzcd}$$