Lecture 18 - Exp. family and estimation in PGM

Thursday, November 9, 2023 2:31 PM

today: exp. families
estimation for PGM

Exponential family

a (flat / canonical) exponential family on $\mathcal{X}$

is a parametric family of dist. on $\mathcal{X}$ defined by two quantities

I) $h(x)\,d\mu(x)$ → reference measure

reference density, base measure → counting measure (discrete R.V.) → $\sum_x$ pmf
→ Lebesgue " (cts R.V.) → $\int_{\mathcal{X}}$ pdf

II) $T: \mathcal{X} \to \mathbb{R}^p$ called "sufficient statistics" vector
aka. feature vector

members of the family will have pmf / pdf

$$p(x;\eta)\,d\mu(x) = \exp\left(\eta^T T(x) - A(\eta)\right) h(x)\,d\mu(x)$$

defining pieces $(+\Omega_X)$

η: "canonical parameter"

$A(\eta)$: log-normalization or cumulant generating function
log partition fct.

$\mathbb{E}[f] = \int f(x) p(x) dx$

- if Ω_X is discrete, then $p(x;\eta)$ is a pmf
  " " cts.; " " a pdf

✱ want $1 = \int_{\mathcal{X}} p(x;\eta)\,d\mu(x) = \int_{\mathcal{X}} \exp(\eta^T T(x))\, e^{-A(\eta)} h(x)\,d\mu(x)$

[discrete: $\sum_x p(x;\eta)$]

$$\Rightarrow \boxed{A(\eta) \triangleq \log\left(\int_{\mathcal{X}} \exp(\eta^T T(x)) h(x)\,d\mu(x)\right)}$$

$Z(\eta)$

domain $\Omega \triangleq \{\eta \in \mathbb{R}^p \mid A(\eta) < \infty\}$
set of valid canonical parameters

clash of notation:
$\Omega \neq \Omega_X \subseteq \mathcal{X} \subseteq \mathbb{R}^p$

note: $A(\eta)$ is convex in η $\Rightarrow$ Ω is convex

⊛ more generally, consider a reparameterization of a subset of the flat exponential family

$\eta: \Lambda \to \Omega$

⊛ more generally, consider a reparameterization of a subset of the flat exponential family

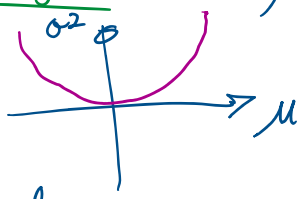
by defining $\eta : \Theta \to \Omega$

(Θ: new set of parameters)

$$p(x;\theta) \triangleq p(x;\eta(\theta)) \quad \text{for } \theta \in \Theta$$

(get a "curved exponential family" if $\eta(\Theta)$ is curved manifold in Ω)

↳ e.g. could consider Gaussians where $N(\mu, \mu^2)$



*: note: any single dist. $p(x)$ can be part of an exponential family by using $h(x) = p(x)$

two examples of families not an exp family:
- unif $(0, \theta)$
- mixture of Gaussians (latent variable model)

Example 1: (multinoulli)

$X \sim \text{Mult}(\pi)$ $\quad \mathcal{X} = \{0,1\}^k$

$\Omega_X = \Delta_k \cap \mathcal{X}$ (one hot encodings)

parameter $\pi \in \Delta_k$; suppose $\pi_j > 0 \ \forall j$

$$p(x;\pi) = \prod_{j=1}^{k} \pi_j^{x_j} = \exp\left(\sum_{j=1}^{k} x_j \log \pi_j\right)$$

π: "think as θ"; $x \in \mathcal{X}$

$$= \exp\left(\sum_{j=1}^{k} x_j \log \pi_j - 0\right)$$

we have $\eta_j(\pi) = \log \pi_j$

$T(x) = x$ $\quad \Omega_X \subseteq \mathbb{R}^k$

$d\mu(x) =$ counting measure on $\mathcal{X}$

$h(x) = \mathbb{1}\{x \in \Omega_X\} = \mathbb{1}\{x \in \Delta_k \cap \mathcal{X}\}$

$A(\eta) = 0$??

$\Theta = \text{int}(\Delta_k)$ $\quad A(\eta(\pi)) = 0 \ \forall \pi \in \Theta$

$\Theta \to$ dim $k-1$

$\eta(\Theta) \to$ " "

$\Omega_X \to \mathbb{N}^K$

we do not have a "minimal exp family"

note: here, for any x s.t. $h(x) \neq 0$

$$\sum_{j=1}^{K} T_j(x) = \sum_j x_j = 1$$

affine linear dep. between components of T

$\Rightarrow$ multiple η's give rise to same distribution $\to$ "overparameterization"

eg $(\eta + \mathbb{1}\alpha)^\top T(x) = \eta^\top T(x) + \alpha \underbrace{T(x)^\top \mathbb{1}}_{1}$

↓ not a minimal exp. family

(*) for a multinoulli; a min. exp family

$$T(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix} \qquad \left[\text{"} x_K = 1 - \sum_{j=1}^{K-1} x_j \text{"} \right]$$

$$Z(\eta) = \sum_{x \in \Omega_X} \exp(\eta^\top T(x)) = \sum_{j=1}^{K-1} e^{\eta_j} + 1$$

$$p(x;\eta) = \exp\left(\sum_{j=1}^{K-1} \eta_j x_j - \underbrace{\log\left(1 + \sum_{j=1}^{K-1} e^{\eta_j}\right)}_{A(\eta)} \right)$$

recall: $\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)]$ (valid for $\eta \in \text{int}(\Omega)$)

for multinoulli; $\frac{\partial A}{\partial \eta_j} = \frac{1}{Z(\eta)} e^{\eta_j} = p(\text{"}x = j\text{"} \mid \eta)$

$= \mathbb{E}_{p(x;\eta)}[T(x)]$ as required //

moment matching can be different than MLE in exp. family (with wrong moments)

gamma dist $\Gamma(\alpha, \beta)$ has $T(x) = \begin{bmatrix} \log x \\ x \end{bmatrix}$

so moment matching with $T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$ will yield a different estimate than MLE

15h28

example 2 : 1d Gaussian

$X \sim N(\mu, \sigma^2)$ $\quad \mathcal{X} = \mathbb{R}$ $\quad \Theta = (\mu, \sigma^2)$ "moment parameterization"

$$p(x; (\mu,\sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \exp\left(\underbrace{-\frac{x^2}{2}}_{\eta_2}\underbrace{\left[\frac{1}{\sigma^2}\right]}_{\eta_2} + x\underbrace{\left[\frac{\mu}{\sigma^2}\right]}_{\eta_1} - \overbrace{\left[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right]}^{A(\eta)}\right)$$

$$T(x) = \begin{bmatrix} x \\ -\frac{x^2}{2} \end{bmatrix} \quad \eta(\theta) = \begin{bmatrix} \mu/\sigma^2 \\ 1/\sigma^2 \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$$

$\eta_2 = \frac{1}{\sigma^2}$ = precision > 0 $\quad h(x) = 1$ (but some people use $h(x) = \frac{1}{\sqrt{2\pi}}$ for Gaussian)

$\eta_1 = \eta_2 \cdot \mu$

$\Omega = \{(\eta_1, \eta_2) : \eta_2 > 0, \eta_1 \in \mathbb{R}\}$

[we'll see later: multivariate Gaussian $T(x) = \begin{bmatrix} x \\ -\frac{xx^T}{2} \end{bmatrix}$ $\quad$ "η_1" $= \Lambda\mu = \Sigma^{-1}\mu$, "$\eta_2$" $= \Lambda = \Sigma^{-1}$]

Example 3 : discrete UGM

let $p \in \mathcal{G}(G)$ G is undirected

with $\psi_c(x_c) > 0 \ \forall c, x_c$

$$p(x) = \frac{1}{Z}\prod_{c \in \mathcal{C}} \psi_c(x_c) = \exp\left(\sum_c \log \psi_c(x_c) - \log Z\right)$$

$$= \exp\left(\sum_{c \in \mathcal{C}} \sum_{y_c \in \mathcal{X}_c} \underbrace{\mathbb{1}\{y_c = x_c\}}_{T_{c,y_c}(x)} \underbrace{\log \psi_c(y_c)}_{\eta_{c,y_c}} - \log Z\right)$$

$$T(x) = \left(\vdots \ \mathbb{1}\{x_c = y_c\} \ \vdots\right)_{c \in \mathcal{C},\, y_c \in \mathcal{X}_c} \qquad \mathcal{X}_c = \{(y_i)_{i \in c} : y_i \in \mathcal{X}_i\}$$

$$\eta(\theta) = \left(\vdots \ \log \psi_c(y_c) \ \vdots\right)_{c \in \mathcal{C},\, y_c \in \mathcal{X}_c}$$

[not a minimal representation]

notes: a) Mult(π) is a special case where have complete graph (1 big clique)

notes: a) Mult(π) is a special case where have complete graph (1 big clique)

b) feature perspective: instead of using all possible indicators $\mathbb{1}\{y_c = x_c\}$ you could use a subset or a function of them for a task

for example: suppose x is a sentence
x_i is a word

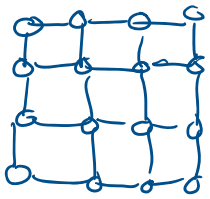
feature on x_i, x_{i+1} e.g. $\mathbb{1}\{x_i$ is a verb, x_{i+1} is a noun$\}$

→ much smaller set of parameters

"parameter sharing"

c) binary Ising model

$x_i \in \{0,1\}$ $|C| \leq 2$



suppose use nodes & pairs (edges) as cliques

⇒ dimension of $T(x)$ $2|V| + 4|E|$ → "overparameterized exp. family"

$\sum_{y_c} T_{c,y_c}(x) = 1$ for any c

⇒ not a min. exp. fam.

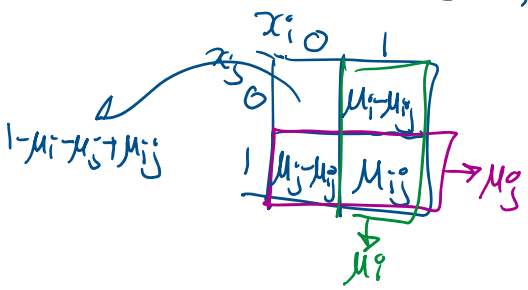
* a minimal representation

is $T(x) = \begin{pmatrix} (x_i)_{i \in V} \\ (x_i x_j)_{\{i,j\} \in E} \end{pmatrix} \begin{matrix} \} n_i \\ \} n_{ij} \end{matrix}$ → dim $|V| + |E|$

$(x_i x_j)_{\{i,j\} \in E} \to \mathbb{1}\{x_i = 1, x_j = 1\}$

$\mathbb{E}\, T(x) = \begin{pmatrix} (\mu_i)_{i \in V} \\ (\mu_{ij})_{\{i,j\} \in E} \end{pmatrix}$

$\mu_{ij} \to p(x_i = 1, x_j = 1)$



properties of A (for general flat exponential family)

properties of A (for general flat exponential family)

- $\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)] \triangleq \mu(\eta)$ "moment vector" (for $\eta \in \text{int}(\Omega)$)
- $\left(\frac{\partial^2 A}{\partial \eta_i \partial \eta_j}\right)_{(i,j)\in[p]^2} = \mathbb{E}_{p(x,\eta)}\left[(T(x)-\mu(\eta))(T(x)-\mu(\eta))^T\right] = \text{Cov}(T(X))$

(proof as exercise)

"cumulant generating fct."

Estimation of parameters DGM / UGM

DGM:

(fully observed)

parametric family $\mathcal{P}_\Theta = \{ p_\theta(x) = \prod_i p(x_i | x_{\pi_i}, \theta_i) : (\theta_1, \dots, \theta_{|V|}) \in \Theta \}$ i.e. no tying of parameters

independent parameterization

$\Theta = \Theta_1 \times \Theta_2 \times \cdots \Theta_{|V|}$

$\Rightarrow$ MLE decouple in $|V|$ independent MLE problems:

$$\{x^{(i)}\}_{i=1}^n \quad p(\text{data}|\theta) = \prod_{i=1}^n p(x^{(i)}|\theta) = \prod_{i=1}^n \prod_{j=1}^{|V|} p(x_j^{(i)} | x_{\pi_j}^{(i)}; \theta_j)$$

$$\log(\quad) = \sum_{j=1}^{|V|} \underbrace{\left(\sum_{i=1}^n \log p(x_j^{(i)} | x_{\pi_j}^{(i)}; \theta_j) \right)}_{\ell_j(\theta_j)}$$

example: for discrete R.V. (multinoulli conditionals) $\Rightarrow \hat{\theta}_j^{MLE}$ = proportion of observations

$$\hat{p}(x_j = k \mid x_{\pi_j} = \text{stuff}) = \frac{\#(x_j = k, x_{\pi_j} = \text{stuff})}{\#(x_{\pi_j} = \text{stuff})}$$

(fully observed DGM is relatively easy often in closed form!)

(*) if have latent variables (ie unobserved variables)

$\Rightarrow$ use E.M. (like mean)

UGM:

example for exp. family

$$p(x;\eta) = \exp\left(\sum_c \langle \eta_c, T_c(x_c) \rangle - A(\eta) \right)$$

$$p(x;\eta) = \exp\left(\sum_c \langle \eta_c, T_c(x_c) \rangle - A(\eta) \right)$$

→ unlike in a DGM, $\log p(x;\eta)$ does not separate $\sum_c f_c(\eta_c)$

gradient ascent on log likelihood:

$$\frac{1}{n} \sum_{i=1}^{n} \log p(x^{(i)} | \eta) = \sum_c \eta_c^T \underbrace{\left[\frac{1}{n} \sum_{i=1}^{n} T_c(x_c^{(i)}) \right]}_{\hat{\mu}_c} - \frac{A(\eta)}{n}$$

$$\nabla_{\eta_c} [\quad] = \hat{\mu}_c - \mu_c(\eta)$$

$$\mu_c(\eta) = \mathbb{E}_{p(x;\eta)}[T_c(x_c)]$$

to compute this, need inference

e.g. Ising model $T_{ij}(x_i, x_j) = x_i x_j$

$$\mathbb{E}[T_{ij}] = \mu_{ij} = p(x_i = 1, x_j = 1 | \eta)$$

here need approximate inference ← sampling [Gibbs sampling] / variational [mean field]