Lecture 20 - exp. family and estimation in PGM

Friday, November 13, 2020 13:34

today: · exp. family
· estimation PGM
· sampling

Exponential family

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

is a parametric family of dist. defined by two quantities:

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

- $h(x)$: reference density
- $d\mu(x)$: base measure ← counting measure (discrete R.V.) / Lebesgue measure (cts. R.V.)

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

members of the family will have dist.

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

- η: "canonical parameter"
- $T(x)$: defining pieces ($\in \Omega_X$)
- $A(\eta)$: log normalization or log partition fct. or cumulant gen. fct.

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

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

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

$\int_{\mathcal{X}} \exp(\eta^T T(x))\, h(x)\, d\mu(x) = Z(\eta)$

$\mathcal{X} \to \Omega_X$

domain $\Omega \triangleq \{\eta \in \mathbb{R}^p \mid A(\eta) < \infty\}$

(set of valid canonical parameters)

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

$\circledast$ more generally, consider a reparameterization of a subset of the family

by defining $\eta : \mathcal{H} \to \Omega$

$\mathcal{H}$: new set of parameters

consider $p(x;\theta) \triangleq p(x;\eta(\theta))$ for $\theta \in \mathcal{H}$

(get a "curved exponential family" if $\eta(\mathcal{H})$ is a curved manifold in Ω)

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

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

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

Example: (Multinomial)

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

$\Omega_{\mathcal{X}} = \Delta_K \cap \mathcal{X}$ (one hot encoding)

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

$$p(x;\pi) = \prod_{j=1}^{K} \pi_j^{x_j} = \exp\Big(\sum_j x_j \log \pi_j\Big)$$

(think as θ)

$$= \exp\Big(\sum_j x_j \log \pi_j - 0\Big)$$

we have $\eta_j(\pi) = \log \pi_j$ $\quad \Omega = \mathbb{R}^K$

$T(x) = x$

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

$h(x) = \mathbb{1}\{x \in \Omega_{\mathcal{X}}\} = \mathbb{1}\{x \in \Delta_K \cap \mathcal{X}\}$

$\mathcal{H} = \text{int}(\Delta_K)$ $\quad A(\eta(\pi)) = 0 \ \forall \pi \in \mathcal{H}$

$\mathcal{H}$	→ dimension	$K-1$
$\eta(\mathcal{H})$	→	$\|$ $K-1$
Ω	→	$\|$ K

we do not have a "minimal exp family"

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

have, $\underbrace{\sum_{j=1}^{k} T_j(x)}_{\text{affine linear dep. between of } T} = \sum_{j=1}^{k} x_j = 1$

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

↳ not a "minimal" exp family

(*) for multinoulli, minimal exp family

$$T(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix}$$

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

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

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

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

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

14h29

example 2: 1d Gaussian

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

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

$$= \exp\left(-\frac{x^2}{2}\left[\frac{1}{\sigma^2}\right] + x\left[\frac{\mu}{\sigma^2}\right] - \left[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right] \right)$$

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

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

$$\eta_2 = \frac{1}{\sigma^2} = \text{precision} > 0$$

$$\eta_1 = \eta_2 \cdot \mu$$

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

[we'll see later: multivariate Gaussian $\Lambda = \Sigma^{-1}$, $\eta = \Lambda\mu = \Sigma^{-1}\mu$]

$$T(x) = \begin{bmatrix} x \\ -\frac{xx^T}{2} \end{bmatrix}$$

Example 3: discrete UGM?

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

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

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

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

$\mathbb{1}\{y_c = x_c\} = T_{c,y_c}(x)$, $\log \psi_c(y_c) = \eta_{c,y_c}$

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

$$\eta(\theta) = \left(\vdots \; \log \psi_c(y_c) \; \vdots \right)$$

[not a minimal representation]

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

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

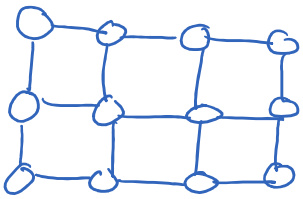
for example: suppose x is sentence, x_i is a word

feature on x_i $\{x_{(i)}$ e.g. $\mathbb{1}\{x_i$ is a verb$\}$.

$x_{i,k}$ is a noun

c) **binary Ising model**

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



suppose use nodes & pairs (edges) as cliques

$\Rightarrow$ dimension of $T(x)$ is $2|V| + 4|E|$ $\rightsquigarrow$ "over parameterized" exp. family

$\sum_{y_c} T_{c,y_c}(x) = 1$ for any c $\Rightarrow$ not a min. exp family

* a minimal representation:

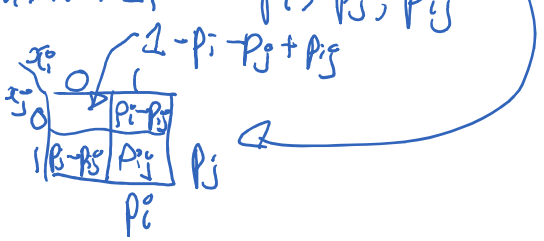
is

$$T(x) = \begin{pmatrix} (x_i)_{i \in V} \\ (x_i x_j)_{\{i,j\} \in E} \end{pmatrix} \rightarrow \text{dim} : |V| + |E|$$

$\mathbb{1}\{x_i = 1, x_j = 1\}$

3 numbers are sufficient

p_i, p_j, p_{ij}



properties of $A(\eta)$

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

(proof as exercise)

Estimation of parameters in DGM

DGM: parametric family $P_{\mathcal{H}} = \{p_\theta(x) = \prod_i p(x_i | x_{\pi_i}, \theta_i)\}$

$\theta = (\theta_1, \dots, \theta_{|V|})$

$\theta \in \mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 \times \dots \mathcal{H}_{|V|}$

independent parameterization

i.e. no tying of parameters

$\Rightarrow$ MLE decouples in $|V|$ independent MLE problems

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

$$\log[\ ] = \sum_{j=1}^{|V|} \underbrace{\left(\sum_{i=1}^{n} \log p(x_j^{(i)} \mid x_{\pi_j}^{(i)}, \theta_j)\right)}_{f_j(\theta_j)}$$

Example: for discrete R.V. $\Rightarrow \theta_j^{MLE}$ = proportion of observations

$$\frac{\#(x_j = k,\ x_{\pi_j} = \text{something})}{\#(x_{\pi_j} = \text{something})}$$

(fully observed DGM is relatively easy)

⊕ if have latent variable (i.e. unobserved variables)

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

UGM:

example for exp. family

$$p(x \mid \eta) = \exp\left(\sum_c \eta_c T_c(x_c) - A(\eta)\right)$$

$\rightsquigarrow$ unlike in a DGM, $\log p(x \mid \eta)$ does not separate as $\sum_c f_c(\eta_c)$

gradient ascent on log likelihood

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

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

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

to compute this, need inference

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

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

here need approx. inference

- sampling
- variational method