Lecture 16 - scribbles

Tuesday, October 25, 2016
14:35

today: • duality (cont.)
• exp. family

dual problem for Max. Ent:

$$(P)\quad \min_q \sum_x q(x)\log\frac{q(x)}{u(x)} \;\Big]\; 1/|\mathcal{X}|$$
$$\left.\begin{array}{l} q(x)\geq 0 \\ \sum_x q(x)=1 \end{array}\right\}\Delta_{|\mathcal{X}|}$$
$$\sum_x q(x)T_j(x)=\alpha_j$$
$$\Big\}\in M$$

$$\mathcal{L}(q,\nu,c)=\sum_x q(x)\log\frac{q(x)}{u(x)}+\sum_j \nu_j\Big(\alpha_j-\sum_x q(x)T_j(x)\Big)+c\Big(1-\sum_x q(x)\Big)$$

$$\Rightarrow\quad q^*(x)=u(x)\exp\big(\nu^T T(x)+c-1\big)$$

exponential family

dual function:

$$g(\nu,c)=\inf_q \mathcal{L}(q,\nu,c)=\mathbb{E}_{q^*}\big[\nu^T T(x)+c-1\big]+\nu^T\alpha-\mathbb{E}_{q^*}\big[\nu^T T(x)\big]$$
$$c-\mathbb{E}_{q^*}[c]$$
$$=\nu^T\alpha+c-\sum_x u(x)\exp(\nu^T T(x))\,e^{c-1}$$

maximize with respect to c: $\nabla_c \rightarrow$
$$1-\underbrace{\sum_x u(x)\exp(\nu^T T(x))}_{\triangleq Z(\nu)}e^{c-1}=0$$
$$e^{c^*-1}=\frac{1}{Z(\nu)}$$

plug back, we get $\max_c g(\nu,c)=\nu^T\alpha-\ln Z(\nu)\triangleq\tilde{g}(\nu)$

• if $\alpha=\frac{1}{n}\sum_{i=1}^n T(x_i)=\mathbb{E}_{\hat{p}_n}[T(x)]$

then $\tilde{g}(\nu)=\frac{1}{n}\sum_{i=1}^n\big[\nu^T T(x_i)-\ln Z(\nu)\big]$

$\ln p(x_i|\nu)$ where $p(x|\nu) \triangleq \exp(\nu^T T(x) - \ln Z(\nu))$

dual problem was $\max_\nu \tilde{g}(\nu) = \max_\nu \log p(x_1, \dots, x_n|\nu)$ i.e. MLE

so ML in the exponential family with $T(x)$ as suff. statistic is equivalent dual of Max Entropy with moment constraint $\mathbb{E}_{\hat{p}_n}[T(x)] = \alpha$

note: $\nabla_\nu \ln Z(\nu) = \frac{1}{Z(\nu)} \nabla_\nu \sum_x u(x) \exp(\nu^T T(x))$

$= \frac{1}{Z(\nu)} \sum_x u(x) \exp(\nu^T T(x)) T(x)$

$p(x|\nu)$

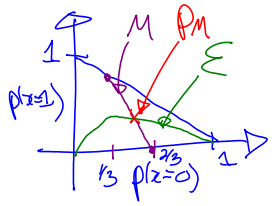
$\nabla_\nu \ln Z(\nu) = \mathbb{E}_{p(x|\nu)}[T(x)] \triangleq \mu(\nu)$ [model moment]

so $\nabla_\nu \tilde{g}(\nu) = \overbrace{\mathbb{E}_{\hat{p}_n}[T(x)]}^{\alpha} - \mu(\nu)$ $\quad T(x) \in \mathbb{R}^1$

$\nabla_\nu \tilde{g}(\nu) = 0 \Rightarrow \mathbb{E}_{p(x|\nu)}[T(x)] = \mathbb{E}_{\hat{p}_n}[T(x)]$ moment matching?

MLE in exponential family $\Leftrightarrow$ moment matching in exponential family

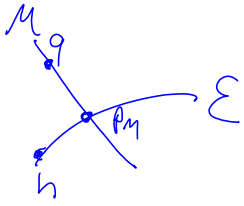
geometry:



$T(x) = \begin{cases} 1 & \text{if } x=0 \\ 1/2 & \text{if } x=1 \\ 0 & \text{if } x=2 \end{cases}$ $\quad A = \{0,1,2\}$

$\mathbb{E}_p[T(x)] = \frac{2}{3} = \alpha$

$\mathcal{E} = \{ p : p(x) = a(x) \exp(\eta^T T(x) - A(\eta)) \}$



for any $q \in M$ and $h \in \mathcal{E}$

$$KL(q \| h) = KL(q \| p_M) + KL(p_M \| h)$$

$KL(q \| h) = E_q[\log \frac{q}{h}]$ "Information Pythagorean thm."

restate our duality result: $p_M = \underset{q \in M}{\operatorname{argmin}} KL(q \| u)$ [Max Ent]

"I-projection" I → information

$p_M = \underset{q \in \mathcal{E}}{\operatorname{argmin}} KL(\hat{p}_n \| q)$ MLE in exp family

"M-projection" M → moment

Exponential family

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

is a parametric family of distributions defined by two quantities

I) $h(x)\, d\mu(x)$ → reference measure on $\mathcal{X}$

reference density · base measure — counting (discrete) / Lebesgue (cts. space)

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

members of family have dist.:

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

defining pieces (Ω_X)

"canonical parameter" · log-normalizer or cumulant gen. function / log partition fct.

if Ω_X is discrete, $p(x;\eta)$ is pmf $\exp(\eta^T T(x) - A(\eta))\, h(x)$

if Ω_X is cts. $p(x;\eta)$ is pdf " "

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

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

$Z(\eta)$

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

* more generally, consider reparameterization as a subset of the family by defining mapping

$\eta : \Theta \to \Omega$ (parameter)

consider $p(x;\Theta) \triangleq p(x;\eta(\Theta))$ for $\Theta \in \Theta$

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

two examples not an exponential family: • Unif $[0,\theta]$
• mixture model

but most standard distributions are (Poisson, Gamma, gaussian, etc...)

Example: (multinomial)

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

$\Omega_{\mathcal{X}} = \Delta_k \cap \mathcal{X}$

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

$$p(x;\pi) = \prod_{j=1}^{k} \pi_j^{x_j} = \exp\left(\sum_{j=1}^{k} x_j (\log \pi_j)\right)$$
$$= \exp(\eta(\pi)^T x - 0)$$

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

$T(x) = x$

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

$h(x) = \mathbb{1}\{x \in \Omega_{\mathcal{X}}\} = \mathbb{1}\{x \text{ has exactly one entry equal to } 1\}$

here $A(\eta(\pi)) = 0$ here $\Theta = \text{int}(\Delta_k)$

$A(\eta(\theta)) = 0\ \forall \theta \in \Theta$

remark: $\Theta \to$ dimension $k-1$

$\mathcal{L} \hookrightarrow \mathbb{R}^K$

here, for any x s.t. $h(x) \neq 0$, $\sum_{j=1}^{K} T_j(x) = 1$

i.e. $\sum_{j=1}^{K} T_j(x) - 1 = 0$

affine linear dep. between components of T

multiple η's map to same distribution "over parameterization"

↳ not a minimal exp. family

for multinomial, minimal exp family $T(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix}$

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

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

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

for multinomial: $\frac{\partial}{\partial \eta_j} A(\eta) = \frac{1}{Z(\eta)} \exp(\eta_j) = p(x=j \mid \eta) = \mathbb{E}_{p(x;\eta)}[T_j(x)]$ as required

⊛ $T(x)$ is called "sufficient" as in statistics

$T : x \mapsto T(x)$ is 'sufficient' for a parametric model P_θ

iff $\forall \theta \in \Theta \quad p_\theta(x) = h(x)\, g(T(x), \theta)$

(functions for family)

i.e. dependence on θ only happens through $T(x)$

iid. model, $\log p(x_1, \dots, x_n \mid \eta) = \sum_{i=1}^{n} \left[\log h(x_i) + \eta^T T(x_i) - A(\eta) \right]$

→ exp. family with reference measure $\tilde{h}(x) = \prod_i h(x_i)$

sufficient $\tilde{T}(x) = \sum_{i=1}^{n} T(x_i)$

log-partition $\tilde{A}(\eta) = n A(\eta)$

Example 2 : Gaussian

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

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

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

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

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

$\eta_1 = \eta_2 \cdot \mu$ $\qquad \Omega = \{(\eta_1, \eta_2) : \eta_2 > 0\}$