Lecture 17 - scribbles

Tuesday, November 6, 2018 14:19

today: - finish MaxENT duality
- exponential family

dual problem for max. entropy

$KL(q \| u)$ $\quad u(x) = \frac{1}{|\mathcal{X}|}$

MaxENT primal form

$$(P) \quad \begin{cases} \min\limits_{q} \sum\limits_{x} q(x) \log \frac{q(x)}{u(x)} \\ q(x) \geq 0 \ \forall x \\ \sum\limits_{x} q(x) = 1 \\ \sum\limits_{x} q(x) T_j(x) = \alpha_j \ \forall j \end{cases}$$

absorb this constraint in domain definition of $KL(q\|u)$ i.e.

$$KL(q\|u) = \begin{cases} +\infty & \text{if } q(x)<0 \text{ for some } x \\ KL(q\|u) & \text{o.w.} \end{cases}$$

$\Delta_{|\mathcal{X}|}$: probability simplex

M : moment matching constraints

$T_j(x)$: feature fct.

$$\mathcal{L}(q, v, c) = \sum_x q(x) \log \frac{q(x)}{u(x)} + \sum_j v_j \left(\alpha_j - \mathbb{E}_q[T_j(x)]\right) + c\left(1 - \sum_x q(x)\right)$$

$$\frac{\partial}{\partial q(x)} = 1 + \log \frac{q(x)}{u(x)} - \sum_j v_j T_j(x) - c = 0$$

$$\Rightarrow \boxed{q^*(x) = u(x) \exp\left(v^T T(x) + c - 1\right)}$$

$\sum_j v_j T_j(x)$

exponential family?

dual function.

exponential family:

dual function:

plug back in $\mathcal{L}(\dots)$ $\quad g(\nu, c) = \mathcal{L}(q^*, \nu, c)$

$$= \mathbb{E}_{q^*}[\nu^T T(x) + c - 1] + \nu^T \alpha - \mathbb{E}_{q^*}[\nu^T T(x)] + c - \mathbb{E}_{q^*}[c]$$

$$= \nu^T \alpha + c - \underbrace{\sum_x u(x) \exp(\nu^T T(x))}_{\triangleq Z(\nu)} \exp(c-1) \qquad \mathbb{E}_{q^*}[1]$$

max $g(\nu, c)$

with respect to c $\quad \nabla_c = 0 \Rightarrow 1 - Z(\nu)\exp(c-1) = 0$

$$\Rightarrow \exp(c^* - 1) = \frac{1}{Z(\nu)}$$

plug back c^* : $\quad \max_c g(\nu, c) = \nu^T \alpha + \underbrace{c^* - Z(\nu)\frac{1}{Z(\nu)}}$

$$\tilde{g}(\nu) \triangleq \nu^T \alpha - \log Z(\nu)$$

If $\alpha = \frac{1}{n}\sum_i T(x_i) = \mathbb{E}_{\hat{p}_n}[T(x)]$

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

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

dual problem is $\max_{\nu} \tilde{g}(\nu) = \max_{\nu} \frac{1}{n} \log p(x_{1:n}|\nu)$ i.e. MLE //

to summarize: ML in the exp. family with $T(x)$ as sufficient statistics is equivalent to max entropy with moment constraints on $T(x)$ where $\alpha = \mathbb{E}_{\hat{p}_n}[T(x)]$

they are Lagrangian dual of each other

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

note: $\nabla_\nu \ln Z(\nu) = \frac{1}{Z(\nu)} \nabla_\nu \sum_x u(x) \exp(\nu^T T(x)) = \sum_x \left(\underbrace{\frac{1}{Z(\nu)} u(x) \exp(\nu^T T(x))}_{p(x|\nu)} T(x) \right)$

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

$\nabla_\nu \tilde{g}(\nu) = \underbrace{\mathbb{E}_{\hat{p}_n}[T(x)]}_{\hat{\mu}_n \text{ "empirical moment"}} - \mu(\nu)$

$\nabla_\nu \tilde{g}(\nu) = 0 \Rightarrow \boxed{\mu(\nu^*) = \hat{\mu}_n}$ i.e. moment matching !

15h32

(see end of old lecture 16 for "KL Pythagorean theorem" and I-projection vs. M-projection for KL + geometry)

Exponential family

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

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 $\mathcal{X}$

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

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

members of 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"

$A(\eta)$: log normalizer or cumulant generating fct. or log partition fct.

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

" " cts, " " " a pdf

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

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

$Z(\eta)$

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

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

(set of valid canonical parameters)

note: $A(\eta)$ is convex in η

$\Rightarrow \Omega$ is convex

(*) more generally, consider a reparameterization of a subset of the family by defining the mapping $\eta : \Theta \to \Omega$

Θ: new set of parameters

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

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

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

* note: any single distribution $p(x)$ can be put in exponential by using $h(x) \triangleq p(x)$

* two examples of family not an exponential family:
  - mixture of Gaussians
  - Unif$(0,\theta)$

Example: (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 $\pi_i > 0 \ \forall i$

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

(π: think as "θ")

$$= \exp\left(\eta(\pi)^T x - 0\right)$$

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

$T(x) = x$

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

$h(x) = \mathbb{1}\{x \in \Omega_X\} = \mathbb{1}\{x$ has exactly one entry equal to $1\}$

$\mathcal{H} = \text{int}(\Delta_K)$

here, $A(\eta(\pi)) = 0 \quad \forall \pi \in \mathcal{H}$

but here $\Omega = \mathbb{R}^K$ (domain; not to be confused with Ω_X sample space)

$\mathcal{H} \to$ dimension $K-1$

$\eta(\mathcal{H}) \to$ " $K-1$

$\Omega \to$ dimension K

here, we do not have "minimal exponential family"

Here, we do not have "minimal exponential family"

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

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

affine linear dep. between components of T

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

↳ not a minimal exp. family

(*) for multinoulli, minimal exp family

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

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

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

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

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

$\log \mathcal{L}(\eta)$

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

$T: x \mapsto T(x)$ is "sufficient" for a parametric family P_θ

iff $\forall \theta \in \Theta \quad p_\theta(x) = \underbrace{h(x)}_{\text{fixed for family}} g(T(x); \theta)$

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

i.i.d. model in exp. family

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

↳ exp. family on $x_{1:n}$, with reference density $\tilde{h}(x) = \prod_i h(x_i)$

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

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

Example 2 : (1D) Gaussian

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

$$p(x; (\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \theta = (\mu, \sigma^2) \quad \text{"moment parameterization"}$$

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

$T(x) = \begin{bmatrix} \quad \end{bmatrix} \qquad \eta = \begin{bmatrix} \quad \end{bmatrix}$

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

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

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