Class notes

Tuesday, February 6, 2018 11:50 AM

$$p(x;\theta) = \exp(\theta^T T(x) - A(\theta))\, h(x)$$

θ: param; $T(x)$: suff stat; $A(\theta)$: normalizer; $h(x)$ → base measure

$\exp(f(x))$

Ex:

exponential Dist.

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{other} \end{cases}$$

$$p(x;\theta) = \exp(-\lambda x + \log \lambda)\, \mathbb{1}(x > 0)$$

$-\lambda$: θ; $\log\lambda$: $-A(\theta)$; $\mathbb{1}(x>0)$: $h(x)$

$h(x) = \mathbb{1}(x > 0)$ $\qquad \theta = -\lambda$

$T(x) = x$

$$A(\theta) = \log\left(\frac{1}{\lambda}\right) = \log\left(-\frac{1}{\theta}\right)$$

Ex: Categorical Dist $\quad x \in \{1, \dots, k\}$ $\quad \mu \in \Delta^K$, $\mu_k = p(x = k)$

$$p(x;\mu) = \prod_{k=1}^{K} \mu_k^{\mathbb{1}(x=k)}$$

$$= \prod_{k=1}^{K} \exp(\log \mu_k\, \mathbb{1}(x=k))$$

$$= \exp\left(\sum_{k=1}^{K} (\log \mu_k)\, \mathbb{1}(x=k)\right)$$

$$\left\langle \begin{bmatrix} \log \mu_1 \\ \vdots \\ \log \mu_K \end{bmatrix}, \begin{bmatrix} \mathbb{1}(x=1) \\ \vdots \\ \mathbb{1}(x=K) \end{bmatrix} \right\rangle \qquad \log\mu, \quad \mathbb{1}_{k-1}(x)$$

$$= \exp\left((\log \mu)_{k-1}^T\, \mathbb{1}_{k-1}(x) + \left(1 - \sum_{k=1}^{K-1} \mathbb{1}(x=k)\right)\left(1 - \sum_{k=1}^{K-1} \log \mu_k\right) \right)$$

$\sum_{j=1}^{K-1} \log \mu_j\, \mathbb{1}(x=j)$ $\qquad \log(1 - \sum \mu_k)$ $\qquad \log \mu_K$

$$= \exp\left(\sum_{j=1}^{K-1} \mathbb{1}(x=j) \log \frac{\mu_j}{\mu_K} + \left(1 - \sum_{j=1}^{K} \log \mu_j\right) \right)$$

inner product; $\log(1 - \sum \mu_k)$

$$T(x) = \begin{bmatrix} \mathbb{1}(x=1) \\ \vdots \\ \mathbb{1}(x=k-1) \end{bmatrix} \qquad \theta = \begin{bmatrix} \log \frac{\mu_1}{\mu_K} \\ \log \frac{\mu_2}{\mu_K} \\ \vdots \\ \log \frac{\mu_{K-1}}{\mu_K} \end{bmatrix}$$

$A(\theta) =$

$$\exp(\theta(\eta)^T T(x) - A(\theta(\eta)))\, h(x)$$

linear form a non-l. x; only θ; only func of x

θ:

$\theta(\eta)$ curved exp

$$p(x;\theta) = \exp(\theta^T T(x) - A(\theta))\, h(x)$$

$$= \frac{1}{\exp(A(\theta))} \exp(\theta^T T(x))\, h(x)$$

$$\nabla_\theta \int \exp(\theta^T T(x))\, h(x)\, dx = \nabla_\theta e^{A(\theta)}$$

$$\int T(x) \exp(\theta^T T(x))\, h(x)\, dx = \nabla_\theta A(\theta)\, e^{A(\theta)}$$

$$\int T(x) \frac{\exp(\theta^T T(x))}{e^{A(\theta)}} h(x)\, dx = \nabla_\theta A(\theta)$$

$$\mathbb{E}[T(x)] = \nabla_\theta A(\theta)$$

$$P(\theta, x | t) = P(x | t, \theta)\, P(\theta | t)$$

x — t — θ

Freq: $P(x|t,\theta) = P(x|t)$

$$= P(x|t)\, P(\theta|t) \qquad \theta \perp\!\!\!\perp x \mid t$$

$$P(\theta | t, x) = \frac{P(\theta, x | t)}{P(x|t)} = P(\theta | t)$$

$T(x)$, θ ↔ η

$A(\theta)$ $\qquad \nabla_\theta A(\theta) = \mathbb{E}[T(x)]$ (μ)

$$\theta = (\nabla_\theta A)^{-1}(\mu)$$

$$P(x|\theta) = h(x) \exp(\theta^T T(x) - A(\theta))$$

$$P(\theta; \tau_0, \nu_0) = \exp(\nu_0 \tau_0^T T(\theta) - B(\nu_0, \tau_0))\, \tilde{h}(\theta)$$

$\exp(\theta\eta^T T(\theta) - A(\eta))\, h(\theta)$

$$T(\theta) = \begin{bmatrix} \theta \\ \nu_0 A(\theta) \end{bmatrix} \qquad B(\cdot) = A(\theta)$$

$\exp(\nu_0 \theta^T \tau_0 - \nu_0 A(\theta))$

$\theta^T($

$$P(\theta | D) \propto P(D, \theta) = P(D | \theta)\, P(\theta)$$

$\{x_1, \dots, x_n\} | \theta)\, P(\theta)$ $\qquad \theta \in \mathbb{R}^d \quad \tilde{T}(\theta) \in \mathbb{R}^{d+1}$

$$\propto \left(\prod_{i=1}^{n} h(x_i)\right) \exp\left(\theta^T \sum_{i=1}^{n} T(x_i) - n A(\theta)\right) \times$$

$$\exp(\nu_0 \theta^T \tau_0 - \nu_0 A(\theta)) \qquad T(\theta) = \begin{bmatrix} \theta \\ A(\theta) \end{bmatrix}$$

$$\propto \quad \exp\left(\theta^T\left(\sum_{i=1}^{n} T(x_i) + \nu_0 \tau_0\right) - A(\theta)(n + \nu_0)\right)$$

$$\exp\left(\theta^T\left(\frac{\sum_{i=1}^{n} T(x_i) + \nu_0 \tau_0}{n + \nu_0}\right)(n + \nu_0) - A(\theta)(\nu_0 + n)\right)$$

τ_0

y

$$p(y;\theta) = \exp[\theta y - A(\theta)]\, h(y)$$

$T(y) = y$

$$P(y; \theta, \sigma^2) = \exp\left[\frac{\theta y - A(\theta)}{\sigma^2}\right] h(y, \sigma^2)$$

σ^2: dispers

$$\mu = \mathbb{E}[y] = A'(\theta)$$

(x_i, y_i) $\qquad x_i \in \mathbb{R}^d$, y_i $\qquad$ (graph: x → y)

$w^T x_i \xrightarrow{g^{-1}} \mu$, $\mu \xrightarrow{g} w^T x_i$

$$g^{-1}(w^T x_i) = \mu \qquad g(\mu) \quad w^T x$$

Param of linear funct w, x_i → η $\xrightarrow{g^{-1}}$ μ $\xrightarrow{(A')^{-1}}$ θ; $\theta \xrightarrow{A'} \mu$; $\mu \xrightarrow{g} \eta$ (g invertible)

$\eta = w^T x$

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

$A'(\theta) = \mathbb{E}[y]$

$$g \equiv (A')^{-1}$$

$$\max_w \log P(D; w) = \sum_{i=1}^{n} \log p(y_i; w)$$

$\mathcal{L}$ $\qquad (x_i, y_i)$ $\qquad$ (graph: x → y) $P(y|x)$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} \theta_i y_i - A(\theta_i)$$

l_i; index by i b/c func x_i

$$\frac{\partial l_i}{\partial w} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta} \frac{\partial \eta}{\partial w}$$

x_i

$$(y_i - A'(\theta)) \frac{\partial \theta_i}{\partial \eta} x_i$$

μ_i

$$\nabla_w \mathcal{L} = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \theta_i}{\partial \eta} x_i\right]$$