Exponential Families

Machine Learning 2

Email: t.p.pandeva@uva.nl

Recap

The distribution set

$\{ p(x|\theta) \mid \theta \in \Theta \}$ is an Exponential Family

$$\Longleftrightarrow$$

$$p(x|\theta) = h(x) \exp[\eta^T T(x) - A(\eta)]$$

$h(x)$ base measure $h(x) \in \mathbb{R}$

$T(x)$ sufficient statistics $T(x) \in \mathbb{R}^d$

η natural parameter $\eta(\theta) \in \mathbb{R}^d$

$A(\eta)$ log partition function

Sufficient statistics

$$x_1, \dots, x_n \overset{iid}{\sim} p(x \mid \theta) \quad \text{(Exp. family)}$$

Likelihood:

$$p(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} \left[h(x_i) \exp\left(y^T T(x_i) - A(y) \right) \right]$$

$$= \underbrace{\left[\prod_{i=1}^{n} h(x_i) \right]}_{} \underbrace{\exp\left(\left(\sum T(x_i) \right)^T y - n A(y) \right)}_{\text{It depends on theta}}$$

Consider θ unknown

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\arg\max}\ p(x_1, \dots x_n \mid \theta) =$$

$$= \underset{\theta \in \Theta}{\arg\max} \left[\underbrace{\sum T(x_i)}_{T_n} \right]^T y - n A(y) \Big)$$

Exponential family recipe

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

1. Identify θ, Θ

2. Find the support $S = \{x \mid p(x|\theta) > 0 \ \forall \theta\}$

3. Write down the $\log p(x|\theta)$

$$\log p(x|\theta) = \boxed{\underbrace{\eta_1(\theta) T_1(x) + \cdots + \eta_s(\theta) \cdot T_s(x)}_{\text{linear in } \eta\,!!}} + B(x) + C(\theta)$$

4. $h(x) = \mathbb{1}_S(x)\exp(B(x))$

5. Express θ as a function of η $(\theta(\eta))$

$$A(\eta) = -C(\theta(\eta))$$

(i) $\mathrm{Pois}(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

1) λ, $\lambda \in \mathbb{R}^+$

2) $k \in \mathbb{N}_0$, $\boxed{S = \mathbb{N}_0}$

3) $\log p(k|\lambda) = \log \frac{\lambda^k e^{-\lambda}}{k!} =$

$= \log \lambda^k + \log e^{-\lambda} - \log k!$

$= \underbrace{\boxed{k \log \lambda}}_{T(k)\,y} \underbrace{- \lambda}_{C(\lambda)} \boxed{\underbrace{- \log k!}_{B(k)}}$

$T(k) = k$ $\qquad \mathbb{E}[T(k)], \mathrm{Var}(T(x))$

$\boxed{y = \log \lambda} \Leftrightarrow \lambda = e^y$

4) $h(k) = \mathbb{1}_S(k) \exp(B(k)) =$

$= \mathbb{1}_{\mathbb{N}_0}(k) \frac{1}{k!}$

$$A(y) = -C(x(y))$$

$$= -(-x(y)) = x(y)$$

$$= e^{y}$$

ii) $\mathrm{Geom}(x|\pi) = (1-\pi)^{x-1}\pi$

1) $\pi \in (0, 1]$

2) $x \in \mathbb{N} =: S$

3) $\log p(x|\pi) = \log\left[(1-\pi)^{x-1}\pi\right]$

$= (x-1)\log(1-\pi) + \log\pi$

$= \underbrace{x \log(1-\pi)}_{T(x)\,y} \underbrace{\boxed{-\log(1-\pi) + \log\pi}}_{C(\pi)}, \quad B(x) = 0$

$T(x) = x$

$y = \log(1-\pi) \iff \boxed{\pi = 1-\exp(y)}$

5) $h(x) = \mathbb{1}_S(x)\exp(\underset{\overset{\shortparallel}{0}}{B(x)}) =$

$= \mathbb{1}_{\mathbb{N}}(x)$

$$A(y) = -C(\pi(y)) =$$

$$= -y - \log(1 - \exp(y))$$

iii) $\text{Cauchy}(x \mid \gamma, \mu) = \frac{1}{\pi \gamma} \frac{1}{1 + \left(\frac{x-\mu}{\gamma}\right)^2}$

1) $\gamma > 0, \ \mu \in \mathbb{R}$

2) $x \in \mathbb{R} = S$

3) $\log p(x \mid \gamma, \mu) =$

$$= -\log(\pi) - \log \gamma \underbrace{- \log\left(1 + \left(\frac{x-\mu}{\gamma}\right)^2\right)}_{T(x)^T \eta}$$

Fix $\gamma = 1$

If $p(x \mid \gamma, \mu)$ is an Exp. family

$\Rightarrow \exists \ T, \eta$ s.t.

$$T(x)^T \eta(\mu) = -\log(1 + (x-\mu)^2) \ \forall \mu$$

$$\underline{[T(x)^T y(\mu) = -\log(1+(x-\mu)^2) \ \forall \mu}$$

$$\mu = \mu_0$$

$$T(x) = \underbrace{\frac{1}{y(\mu_0)}}_{\text{const}} [-\log(1+(x-\mu_0)^2)$$

$$y(\mu) = \frac{-\log(1+(x-\mu)^2)}{\frac{1}{y(\mu_0)}[-\log(1+(x-\mu_0)^2)}$$

$y(\mu)$ depends on x $\Rightarrow$ contradiction to the def. of Exp family

iv) $\text{Log Normal}(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{[\ln x - \mu]^2}{2\sigma^2}\right)$

1) $\mu \in \mathbb{R}, \sigma > 0$

2) $x > 0, S = \mathbb{R}^+$

3) $\log p(x \mid \mu, \sigma) =$

$y(\sigma, \mu) = (y_1, y_2)$

$T(x) = (T_1(x), T_2(x))$

$$-\log(x) - \log(\sigma) - \log(\sqrt{2\pi}) - \frac{[\log(x)]^2}{2\sigma^2} + \frac{\log(x)\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} =$$

$T_1(x)\, y_1$

$T_2(x)\, y_2$

$$= \underbrace{-\frac{[\log(x)]^2}{2\sigma^2} + \frac{\log(x)\mu}{\sigma^2}}_{T(x)^T y} + \underbrace{\log\sigma - \frac{\mu^2}{2\sigma^2}}_{c(\sigma, \mu)} - \underbrace{\log(x) - \log(\pi)}_{B(x)}$$

$$T_1(x) = \log(x)$$

$$T_2(x) = (\log(x))^2$$

$E[T_2(x)]$

$$y_1 = \frac{\mu}{\sigma^2}$$

$$y_2 = \frac{-1}{2\sigma^2}$$

5) $h(x) = \mathbb{1}_S(x) \exp(B(x)) = \mathbb{1}_{\mathbb{R}^+}(x) \frac{1}{\sqrt{2\pi}x}$

6) $$y_2 = \frac{-1}{2\sigma^2} \Leftrightarrow \sigma^2 = -\frac{1}{2y_2}$$

$$y_1 = \frac{\mu}{\sigma^2} \Leftrightarrow \mu = y_1\sigma^2 = -\frac{y_1}{2y_2}$$

$$A(y) = -\log\left(-\frac{1}{2y_2}\right)^{1/2} + \frac{\left(\frac{y_1}{2y_2}\right)^2}{-\frac{1}{2y_2}} =$$

$$\boxed{= -\frac{1}{2}\log(-2y_2) - \frac{y_1}{4y_2^2}}$$

$$A(y) = -C(\sigma, \mu) = -\left[\log\sigma - \frac{\mu^2}{2\sigma^2}\right]$$

2. $E[T(x)] = \frac{d}{dy} A(y)$

$Cov[T_i(x), T_j(x)] = \frac{d^2}{dy_i dy_j} A(y)$

i) $T(k) = k$ $\quad$ $y = \ln \lambda$

$A(y) = \exp(y)$

$E[T(k)] = \frac{d}{dy} A(y) = \exp(y)$

$k \sim Pois(k|\lambda)$

$= \lambda$

$Var[T(k)] = \frac{d^2}{dy^2} A(y) = \exp(y)$

$= \lambda$

i) $E[k] = \frac{d}{dy} A(y) = \frac{d}{dy} e^y = e^y = \lambda$

$Var[k] = \frac{d^2}{dy^2} A(y) = \frac{d^2}{dy^2} e^y = e^y$
$= \lambda$

ii) $E[X] = \frac{d}{dy} A(y) = \frac{d}{dy} \left[-y - \log(1-\exp(y)) \right]$

$$= \underbrace{1 + \frac{e^y}{1-e^y}}_{} = \cdots = \frac{1}{p}$$

$$Var(X) = \frac{d}{dy}\left(1 + \frac{e^y}{1-e^y}\right) = \frac{e^y}{(1+e^y)^2} = \frac{1-p}{p}$$

iv) $\mathbb{E}[\ln x] = \frac{d}{dy_1} A(y) = \frac{-y_1}{2y_2} = \mu$

$$E[(\ln x)^2] = \frac{d}{dy_2} A(y) = -\frac{1}{2y_2} = \mu^2 + \sigma^2$$

$$\mathrm{var}[\ln x] = \frac{d^2}{dy_1^2} A(y) = -\frac{1}{2y_2} = \sigma^2$$

$$\mathrm{Var}[(\ln x)^2] = \frac{d^2}{dy_2^2} A(y) = \frac{-y_1^2}{2y_2^3} + \frac{1}{2y_2^2}$$

$$= 4\mu^2\sigma^2 + 2\sigma^4$$

$$\mathrm{Cov}[\ln x, (\ln x)^2] = \frac{d^2}{dy_2 dy_1} A(y) = \frac{y_1}{2y_2^2} =$$

$$= 2\mu\sigma^2$$

3 $\quad p(y|\tau, \nu) \propto \exp(\tau^T y - \nu A(y))$

$$\underbrace{p(y|X, \underbrace{\tau, \nu}_{\text{prior parameters}})}_{\text{posterior}} \propto \underbrace{p(X|y)}_{\text{likelihood}} \underbrace{p(y|\tau, \nu)}_{\text{prior}}$$

$$p(y|\tau, \nu) \in \{p(y|\theta) \mid \theta \in \Theta\}$$

$$\Rightarrow p(y|\tau, \nu) \in \{p(y|\theta) \mid \theta \in \Theta\}$$

$$p(y|\tau, \nu) \propto \exp(\tau^T y - \nu A(y))$$

i) $y = \ln \lambda$, $A(y) = \exp(y) = \lambda$

$$p(y|\tau, \nu) \propto \exp(\tau^T y - \nu A(y))$$

$$= \exp(\tau \ln \lambda - \nu \lambda)$$

$$= \lambda^{\tau} e^{-\nu \lambda}$$

$$\boxed{\begin{aligned} \tau &= \alpha - 1 \\ \nu &= \beta \end{aligned}} \quad = \lambda^{\alpha-1} e^{-\beta \lambda}$$

$$\Rightarrow p(y|\tau, \nu) = \frac{1}{\Gamma(\alpha)} \alpha^{\beta} \lambda^{\alpha-1} e^{-\beta \lambda}$$

ii) $\mathrm{Geom}(x|\pi) = (1-\pi)^{x-1}\pi$

$y = \log(1-\pi)$

$A(y) = -y - \log(1-\exp(y))$

$= \log(1-\pi) - \log\pi$

$$p(y|\tau,\nu) \propto \exp(\tau^T y - \nu A(y))$$

$$= \exp(\tau \log(1-\pi) - \nu(\log(1-\pi) - \log\pi))$$

$$= (1-\pi)^{\tau}(1-\pi)^{-\nu}\pi^{\nu}$$

$$= \boxed{(1-\pi)^{\tau-\nu}\pi^{\nu}}$$

$\alpha = \tau - \nu + 1$

$\beta = \nu + 1$

$$\boxed{\mathrm{Beta}(\pi|\alpha,\beta) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)}}$$

$$= (1-\pi)^{\alpha-1}\pi^{\beta-1}$$

iv) $\alpha \equiv 1$ wlog

$$y = \left(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\right), \quad A(y) = \frac{-y_1^2}{4y_2} - \frac{1}{2}\ln(-2y_2)$$

$$\sigma = 1 \Rightarrow y = \Big(\overset{y_1}{\mu}, \overset{y_2}{-\frac{1}{2}}\Big)$$

$$A(y_1) = \frac{\mu^2}{2}$$

$$p(y_1 \mid \tau, \nu) \propto \exp(\tau^T y_1 - \nu A(y_1))$$

$$= \exp\left(-\frac{\nu\mu^2 - 2\tau\mu}{2}\right)$$

$$\boxed{\propto} \exp\left(-\frac{\nu\mu^2 - 2\tau\mu + \frac{\tau^2}{\nu}}{2}\right)$$

$$= \exp\left(-\frac{\nu\left(\mu - \frac{\tau}{\nu}\right)^2}{2}\right)$$

$$= \exp\left(-\frac{v\left(\mu-\frac{\tau}{v}\right)^2}{2}\right)$$

$$\underbrace{v = \frac{1}{\sigma_0^2}, \quad \frac{\tau}{v} = \mu_0}$$

$$\hookrightarrow \underbrace{= \frac{1}{\sqrt{2\pi\sigma_0}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)}$$

$$\Rightarrow P(\mu \mid \sigma = 1, \tau, v) \triangleq \mathcal{N}(\mu_0, \sigma_0^2)$$

4

prior	posterior
τ	$\tau + \sum T(x_i)$
ν	$\nu + n$

$x_1, \ldots, x_n \overset{iid}{\sim} p(X|\theta)$ exp. family

i) $\boxed{\begin{aligned} \tau &= \alpha - 1 \\ \nu &= \beta \end{aligned}}$ $\quad p(y|\tau,\nu) = \text{Gamma}(\alpha,\beta)$

$$\alpha = \underbrace{\tau}+1 \xrightarrow{p \to p} \underbrace{\tau + \sum x_i} + 1$$

$$= \alpha + \sum x_i$$

$$\beta = \nu \longrightarrow \nu + n = \beta + n$$

$$p(y|X,\alpha,\beta) = \text{Gamma}\left(\alpha + \sum x_i, \beta + n\right)$$

ii) $\alpha = \tau - \upsilon + 1$

$\beta = \upsilon + 1$

$$p(y \mid \tau, \upsilon) = \frac{\pi^{\alpha - 1}(1-\pi)^{\beta - 1}}{B(\alpha, \beta)}$$

$$p(y \mid X, \tau, \upsilon) = \frac{\pi^{? - 1}(1-\pi)^{? - 1}}{B(?, ?)}$$

Prior	Posterior
τ	$\tau + \sum T(x_i)$
υ	$\upsilon + n$
α $(\tau - \upsilon + 1)$	$\tau + \sum x_i - (\upsilon + n) + 1$ $= \underbrace{\tau - \upsilon + 1}_{\alpha} + \sum x_i - n$
β $(= \upsilon + 1)$	$\upsilon + n + 1 = \beta + n$

$$P(y \mid X, \alpha, \beta)$$

$$= \mathrm{Beta}\Big(\underbrace{\alpha + \sum x_i - n}_{\alpha_n}, \underbrace{\beta + y}_{\beta_n}\Big)$$

iv)

prior	posterior
τ	$\tau_n = \tau + \sum T(x_i)$
ν	$\nu_n = \nu + n$

$$\sigma_0^2(\tau, \nu) = \frac{1}{\nu}$$

$$\sigma_0^2\left(\tau + T(x_i), \nu + n\right) = \frac{1}{\nu + n} =$$

$$= \frac{1}{\nu}\left(\frac{1}{1 + \frac{n}{\nu}}\right) =$$

$$= \sigma_0^2\left(\frac{1}{1 + \sigma_0^2 n}\right)$$

$$\left(=: \sigma_{0_n}^2 \right)$$

$$\mu_0(\tau, v) = \frac{\tau}{v}$$

$$\mu_{0_n} = \mu_0\left(\tau + \sum T(x_i), v+n\right) =$$

$$= \frac{\tau + \sum x_i}{v+n} =$$

$$= \underbrace{v}_{\frac{1}{\sigma_0^2}}\left(\underbrace{\frac{\tau}{v}}_{\mu_0} + \underbrace{\frac{1 \sum \ln(x_i)}{v}}_{\sigma_0^2}\right)\underbrace{\frac{1}{v+n}}_{\sigma_{0n}^2}$$

$$= \frac{1}{\cancel{\sigma_0^2}}\left(\mu_0 + \sigma_0^2 \sum \ln x_i\right)\frac{\cancel{\sigma_0^2}}{1+\sigma_0^2 n}$$

$$= \frac{\mu_0 + \sigma_0^2 \sum \ln x_i}{1 + \sigma_0^2 n}$$

MLE & MAP for the Exp. family

$$\hat{\theta}_{MLE} = \underset{\theta}{\arg\max}\; p(x_1 \ldots x_n | \theta)$$

$$= \underset{\theta}{\arg\max}\; \log p(x_1 \ldots x_n | \theta)$$

$$= \underset{\theta}{\arg\max} \left(\left(\sum_{i=1}^{n} T(x)\right)^T y(\theta) - n A(y(\theta)) \right)$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\arg\max}\; p(x_1 \ldots x_n | \theta)\, p(\theta | \tau, \nu)$$

$$\ldots = \underset{\theta}{\arg\max} \left(\sum_{i=1}^{n} T(x) + \tau\right)^T y - (\nu + n) A(y)$$

$$= \underset{\theta}{\arg\min}\; -\left(\sum T(x) + \tau\right)^T y + (\nu + n) A(y)$$

$$\text{c)}\quad T(x) = x \qquad A(y) = \lambda$$

$$y = \ln \lambda$$

$$\lambda_{MAP} = \underset{\lambda}{\operatorname{argmin}} \underbrace{-\left(\sum_{i=1}^{n} x_i + \tau\right) \ln \lambda + (\nu+n)\lambda}_{f(\lambda)}$$

$$\frac{\partial}{\partial \lambda} f(\lambda) = -\frac{\left(\sum x_i + \tau\right)}{\lambda} + (\nu+n) \stackrel{!}{=} 0$$

$$\Rightarrow \lambda_{MAP} = \frac{\sum x_i + \tau}{\nu + n} = \frac{\sum x_i + \alpha - 1}{\beta + 1}$$

Conjugate Prior Recipe

- $p(\eta|\tau,\nu) \propto \exp\left(\tau^T \eta - \nu A(\eta)\right)$ (1)

- write down η and $A(\eta)$ as functions of θ (i.e. $\eta(\theta)$, $A(\eta(\theta))$)

  original parametrization

- Plug $\eta(\theta)$ and $A(\eta(\theta))$ into (1)

  $$p(\eta(\theta)|\tau,\nu) \propto \exp\left(\tau^T \eta(\theta) - \nu A(\eta(\theta))\right) \quad (2)$$

- "Guess" the distribution with parameters γ

- write down τ, ν as functions of γ ($\tau(\gamma)$, $\nu(\gamma)$) and plug into (2)