Exponential Families

Leila Wehbe

March 19, 2013

Exponential Families

Functions of the sort:

$$p(x, \theta) = \exp(< \phi(x), \theta > -g(\theta))$$

Where $\phi(x)$ is a sufficient statistic: "no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter"

Exponential Families

$$p(x,\theta) = \exp(< \phi(x), \theta > + g(\theta))$$

$g(\theta) = log \sum_x \exp(< \phi(x), \theta >)$ is the partition function.

$g(\theta) = log \int_x \exp(< \phi(x), \theta >)dx$ if x continuous

$$\begin{aligned}
\int p(x,\theta)dx &= \int \exp(< \phi(x), \theta > - g(\theta))dx \\
&= \frac{1}{\exp(g(\theta))} \int \exp(< \phi(x), \theta >)dx = 1 \\
\exp(g(\theta)) &= \int \exp(< \phi(x), \theta >)dx \\
g(\theta) &= log \int_x \exp(< \phi(x), \theta >)dx
\end{aligned}$$

Example: Bernoulli distribution

$$p(x) = p^x(1-p)^{(1-x)}$$

$$\phi(x) = x$$

$$\theta = \log(\frac{p}{1-p})$$

$$g(\theta) = \log(1 + e^\theta)$$

$p(x = 1) = \frac{e^\theta}{1+e^\theta} = \frac{1}{e^{-\theta}+1}$ and $p(x = 0) = \frac{1}{1+e^\theta}$

Example: Normal distribution

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

$$\phi(x) = [x, x^2]$$

$$\theta = [\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}]$$

$$g(\theta) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2)$$

Log Partition Function generates cumulants

$$
\begin{aligned}
\partial_\theta g(\theta) &= \partial_\theta \log \int \exp < \phi(x), \theta > dx \\
&= \frac{\int \phi(x) \exp < \phi(x), \theta > dx}{\int \exp < \phi(x), \theta > dx} \\
&= \int \phi(x) \exp(< \phi(x), \theta > - g(\theta)) dx \\
&= \mathbf{E}[\phi(x)]
\end{aligned}
$$

Log Partition Function generates cumulants

$$
\begin{aligned}
\partial_\theta^2 g(\theta) &= \partial_\theta \int \phi(x) \exp(< \phi(x), \theta > - g(\theta)) dx \\
&= \int \phi(x)[\phi(x)^\top - \partial_\theta g(\theta)] \exp(< \phi(x), \theta > - g(\theta)) dx \\
&= \mathbf{E}[\phi(x)\phi(x)^\top] - \mathbf{E}[\phi(x)]\mathbf{E}[\phi(x)]^\top
\end{aligned}
$$

Example: Bernoulli distribution

$$g(\theta) = \log(1 + e^{\theta})$$

$$\mathbf{E}[x] = \partial_{\theta} g(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = p(x = 1)$$

$$\mathrm{Var}[x] = \partial_{\theta}^{2} g(\theta) = \frac{e^{\theta}}{[1+e^{\theta}]^{2}} = p(x = 1)p(x = 0)$$

Example: Poisson distribution

$$
\begin{aligned}
p(x) &= \frac{\lambda^x e^{-\lambda}}{x!} \\
\phi(x) &= x \\
g(\theta) &= e^\theta = \lambda \\
p(x) &= \frac{1}{x!}\exp(x\theta - e^\theta) = \frac{[e^\theta]^x e^{-e^\theta}}{x!} = \frac{\lambda^x e^{-\lambda}}{x!}
\end{aligned}
$$

$$\mathbf{E}[x] = \partial_\theta g(\theta) = e^\theta = \lambda$$

$$\mathrm{Var}[x] = \partial_\theta^2 g(\theta) = e^\theta = \lambda$$

MLE

Write down likelihood:
$\log p(X|\theta) = \log \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} < \phi(x_i), \theta > -g(\theta)$

Differentiate: $\partial_\theta \log p(X;\theta) = m[\frac{1}{m}\sum_{i=1}^{n} \phi(x_i) - \mathbf{E}[\phi(x)]]$

$\frac{1}{m}\sum_{i=1}^{n} \phi(x_i)$ is the sample average.

MLE

For a bernoulli distribution:

$\phi(x) = x$ and $g(\theta) = \log(1 + e^{\theta})$

$$\mathbf{E}[x] \quad = \quad \frac{1}{1 + e^{-\theta}} = \frac{\sum_i x_i}{N}$$

Conjugate Priors

Incorporate prior is similar to adding fake data:
$p(\theta) \propto p(X_{\text{fake}}|\theta)$
$p(\theta|X) \propto p(X|\theta)p(X_{\text{fake}}|\theta) = p(X \cup X_{\text{fake}}|\theta)$

$$
\begin{aligned}
p(\theta|\mu_0, m_0, X) &\propto p(\theta|\mu_0, m_0)p(X|\theta) \\
&\propto \exp(< m_0\mu_0 + \sum_{i=1}^{m} \phi(x_i), \theta > -(m_0 + m)g(\theta))
\end{aligned}
$$

Conjugate Priors

The prior is also in the exponential family:

$$
\begin{aligned}
p(\theta|\mu_0, m_0) &= \exp(m_0 < \mu_0, \theta > - m_0 g(\theta) - h(m_0\mu_0, m_0)) \\
&= \exp(< \phi(\theta), \rho > - h(\rho)) \text{ where } \phi(\theta) = (\theta, -g(\theta))
\end{aligned}
$$

MAP with generalized laplace smoothing

$\frac{1}{n+m}\sum_{i=1}^{n}\phi(x_i)+\frac{m}{n+m}\mu_0$

For normal distribution:

MLE: $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$ and $\sigma^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\hat{\mu}^2$

MAP:
$\hat{\mu}=\frac{1}{n+n_0}\sum_{i=1}^{n}x_i+\frac{m}{n+m}\mu$ and $\sigma^2=\frac{1}{n+n_0}\sum_{i=1}^{n}x_i^2+\frac{n_0}{n+n_0}\mathbf{I}-\hat{\mu}^2$

Conjugate Priors