Exercise Sheet #9

Solution

December 20, 2018

1 Famous Exponential Families

To show that a given distribution is part of the exponential family, we need to write in the the canonical form

$$p(x|w) = \exp\left(\phi(x)^\intercal w - \log Z(w)\right) \tag{1}$$

(a) Bernoulli Distribution

$$\begin{aligned}
p(x|f) = f^x(1-f)^{1-x} &= \exp\left(\log(f^x(1-f)^{1-x})\right) = \exp\left(x \cdot \log(f) + (1-x)\log(1-f)\right) \\
&= \exp\left(x \cdot \log(f) - x \cdot \log(1-f) + \log(1-f)\right) \\
&= \exp\left(x \cdot (\log(f) - \log(1-f)) + \log(1-f)\right) \\
&= \exp\left(\underbrace{x}_{:=\phi(x)} \cdot \underbrace{log\left(\frac{f}{1-f}\right)}_{:=w} + \log(1-f)\right) \\
&= \exp\left(\phi(x)^\intercal w + \log(1 - \frac{e^w}{1+e^w}\right) \\
&= \exp\left(\phi(x)^\intercal w + \log(\frac{1}{1+e^w}\right) \\
&= \exp\left(\phi(x)^\intercal w + \log(1) - \log(1+e^w)\right) \\
&= \exp\left(\phi(x)^\intercal w - \log(\underbrace{1+e^w}_{:=Z(w)})\right) \\
&= \exp\left(\phi(x)^\intercal w - \log(Z(w))\right)
\end{aligned}$$

Therefore, for the Bernoulli distribution, we have the **sufficient statistics** $\phi(x) = x$, the **natural parameters** $w = \log\left(\frac{f}{1-f}\right)$ and the **partition function** $Z(w) = 1 + e^w$.

(b) Gamma Distribution

$$
\begin{aligned}
p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} &= \exp\left(\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}\right)\right) \\
&= \exp\left(\alpha\log(\beta) - \log(\Gamma(\alpha)) + (\alpha-1)\log(x) - \beta x\right) \\
&= \exp\left(\underbrace{\begin{bmatrix} x & \log(x)\end{bmatrix}}_{:=\phi(x)^\intercal}\underbrace{\begin{bmatrix} -\beta \\ (\alpha-1)\end{bmatrix}}_{:=w} - \log\left(\frac{\Gamma(\alpha)}{\beta^\alpha}\right)\right) \\
&= \exp\left(\phi(x)^\intercal w - \log\left(\underbrace{\frac{\Gamma(w_2+1)}{-w_1^{w_2+1}}}_{:=Z(w)}\right)\right) \\
&= \exp\left(\phi(x)^\intercal w - \log(Z(w))\right)
\end{aligned}
$$

Therefore, for the Gamma distribution, we have the **sufficient statistics** $\phi(x) = \begin{bmatrix} x \\ \log(x)\end{bmatrix}$, the **natural parameters** $w = \begin{bmatrix} -\beta \\ (\alpha-1)\end{bmatrix}$ and the **partition function** $Z(w) = \frac{\Gamma(w_2+1)}{(-w_1)^{w_2+1}}$.

(c) Dirichlet Distribution

$$
\begin{aligned}
p(x|\alpha) = \frac{1}{B(\alpha)}\prod_{i=1}^n x_i^{\alpha_i-1} &= \exp\left(\log\left(\frac{1}{B(\alpha)}\prod_{i=1}^n x_i^{\alpha_i-1}\right)\right) \\
&= \exp\left(\sum_{i=1}^n(\alpha_i-1)\log(x_i) - \log(B(\alpha))\right) \\
&= \exp\left(\underbrace{\begin{bmatrix}\log(x_1) & \log(x_2) & \dots & \log(x_n)\end{bmatrix}}_{:=\phi(x)^\intercal}\underbrace{\begin{bmatrix}\alpha_1-1 \\ \alpha_2-1 \\ \vdots \\ \alpha_n-1\end{bmatrix}}_{:=w} - \log(\underbrace{B(\alpha)}_{:=Z(w)})\right) \\
&= \exp\left(\phi(x)^\intercal w - \log(Z(w))\right)
\end{aligned}
$$

Therefore, for the Dirichlet distribution, we have the **sufficient statistics** $\phi(x) = \begin{bmatrix}\log(x_1) \\ \log(x_2) \\ \vdots \\ \log(x_n)\end{bmatrix}$, the **natural parameters** $w = \begin{bmatrix}\alpha_1-1 \\ \alpha_2-1 \\ \vdots \\ \alpha_n-1\end{bmatrix}$ and the **partition function** $Z(w) = B(w+1)$.

(d) Multivariate Gaussian Distribution

$$
\begin{aligned}
p(x|\mu,\Sigma) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal\Sigma^{-1}(x-\mu)\right) \\
&= \exp\left(-\frac{1}{2}(x-\mu)^\intercal\Sigma^{-1}(x-\mu) - \log\left((2\pi)^{d/2}|\Sigma|^{1/2}\right)\right) \\
&= \exp\left(-\frac{1}{2}x^\intercal\Sigma^{-1}x + x^\intercal\Sigma^{-1}\mu - \frac{1}{2}\mu^\intercal\Sigma^{-1}\mu - \log\left((2\pi)^{d/2}|\Sigma|^{1/2}\right)\right) \\
&= \exp\left(-\frac{1}{2}x^\intercal\Sigma^{-1}x + x^\intercal\Sigma^{-1}\mu - \log\left((2\pi)^{d/2}|\Sigma|^{1/2}\exp\left(\frac{1}{2}\mu^\intercal\Sigma^{-1}\mu\right)\right)\right) \\
&= \exp\left(trace\left(-\frac{1}{2}\Sigma^{-1}xx^\intercal\right) + x^\intercal\Sigma^{-1}\mu - \log\left((2\pi)^{d/2}|\Sigma|^{1/2}\exp\left(\frac{1}{2}\mu^\intercal\Sigma^{-1}\mu\right)\right)\right) \\
&= \exp\left(\begin{bmatrix}x^\intercal & vec^\intercal(xx^\intercal)\end{bmatrix}\begin{bmatrix}\Sigma^{-1}\mu \\ -\frac{1}{2}\,vec(\Sigma^{-1})\end{bmatrix} - \log\left((2\pi)^{d/2}|\Sigma|^{1/2}\exp\left(\frac{1}{2}\mu^\intercal\Sigma^{-1}\mu\right)\right)\right) \\
&= \exp\left(\underbrace{\begin{bmatrix}x^\intercal & xx^\intercal\end{bmatrix}}_{:=\phi(x)^\intercal}\underbrace{\begin{bmatrix}\Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1}\end{bmatrix}}_{:=w} - \log\left(\underbrace{(2\pi)^{d/2}|-2w_2^{-1}|^{1/2}\exp\left(-\frac{1}{4}w_1^\intercal w_2^{-1}w_1\right)}_{:=Z(w)}\right)\right) \\
&= \exp\left(\phi(x)^\intercal w - \log(Z(w))\right)\,,
\end{aligned}
$$

where we denote with *vec* the vectorization operator that convert a matrix into column vector. This notation is commonly dropped for convenience. Therefore, for the Multivariate Gaussian distribution, we have the **sufficient statistics** $\phi(x) = \begin{bmatrix}x \\ xx^\intercal\end{bmatrix}$, the **natural parameters** $w = \begin{bmatrix}\Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1}\end{bmatrix}$ and the **partition function** $Z(w) = (2\pi)^{d/2}|-2w_2^{-1}|^{1/2}\exp\left(-\frac{1}{4}w_1^\intercal w_2^{-1}w_1\right)$.

2 Maximum Likelihood Inference

We can compute the *maximum likelihood estimate* w_{ML} by solving

$$
\frac{1}{m}\sum_i \phi(x_i) = \nabla_w \log Z(w) \tag{2}
$$

(a) Bernoulli Distribution

For the Bernoulli Distribution we can compute the gradient of the logarithm of the partition function:

$$
\nabla_w \log(Z(w)) = \nabla_w \log\left(1+e^w\right) = \frac{e^w}{1+e^w} \overset{e^w=\frac{f}{1-f}}{=} f
$$

Combining this with the left side of (2), we get the maximum likelihood estimate for the standard parameters

$$
f_{\mathrm{ML}} = \frac{1}{m}\sum_i \phi(x_i) \overset{\phi(x)=x}{=} \frac{1}{m}\sum_i x_i =: \bar{x}.
$$

and in terms of the natural parameters

$$w_{\mathrm{ML}} = \log\left(\frac{\bar{x}}{1-\bar{x}}\right).$$

(b) Multivariate Gaussian Distribution

Let us compute the gradient for the multivariate normal distribution. In the first term, we leave out terms that do not depend on w and use $|A^{-1}| = |A|^{-1}$.

$$\begin{aligned}
\nabla_w \log(Z(w)) &= \nabla_w \left(-\frac{1}{4}w_1^\intercal w_2^{-1} w_1 - \frac{1}{2}\log|-2w_2|\right) \\
&= \begin{bmatrix}\nabla_{w_1} \\ \nabla_{w_2}\end{bmatrix}\left(-\frac{1}{4}w_1^\intercal w_2^{-1} w_1 - \frac{1}{2}\log|-2w_2|\right) \\
&= \begin{bmatrix} -\frac{1}{2}w_2^{-1}w_1 \\ \frac{1}{4}w_2^{-1}w_1 w_1^\intercal w_2^{-1} - \frac{1}{2}w_2^{-1}\end{bmatrix}
\end{aligned}$$

where we have used the matrix derivatives $\nabla_{w_2} w_1^\intercal w_2^{-1} w_1 = -w_2^{-\intercal} w_1 w_1^\intercal w_2^{-\intercal} = -w_2^{-1} w_1 w_1^\intercal w_2^{-1}$ (due to symmetry of w_2) and $\nabla_{w_2} \log|-2w_2| = w_2^{-1}$. Given the sufficient statistics $\phi(x)$ from 1 (d), we get two equations to solve,

$$-\frac{1}{2}w_2^{-1}w_1 = \frac{1}{m}\sum_i x_i =: \bar{x} \tag{i}$$

$$\frac{1}{4}w_2^{-1}w_1 w_1^\intercal w_2^{-1} - \frac{1}{2}w_2^{-1} = \frac{1}{m}\sum_{ij} x_i x_j =: \bar{S} \tag{ii}$$

Plugging (i) into (ii), we get

$$\begin{aligned}
\bar{S} &= \bar{x}\bar{x}^\intercal - \frac{1}{2}w_2^{-1} \\
w_2^{-1} &= 2\left(\bar{x}\bar{x}^\intercal - \bar{S}\right) \\
\Sigma_{\mathrm{ML}} &= \bar{S} - \bar{x}\bar{x}^\intercal
\end{aligned}$$

where we have used from 1 (d) that $w_2 = -\frac{1}{2}\Sigma^{-1}$. Using this result in (i) and the definition of $w_1 = \Sigma^{-1}\mu$, we get

$$\begin{aligned}
\Sigma w_1 &= \bar{x} \\
w_1^{\mathrm{ML}} &= \Sigma_{\mathrm{ML}}^{-1}\bar{x} \\
\Sigma^{-1}\mu &= \Sigma^{-1}\bar{x} \\
\mu_{\mathrm{ML}} &= \bar{x}.
\end{aligned}$$

That means the maximum likelihood estimate for the mean and covariance of a Gaussian is given by the sample mean and covariance.

3 Conjugate Prior Inference

(a)

We re-write the scalar Gaussian distribution in the form of the exponential family:

$$\begin{aligned}
p(x_i|\sigma^2) &= \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \exp\Big(\underbrace{-\frac{(x_i-\mu)^2}{2}}_{=:\phi(x)} \underbrace{\sigma^{-2}}_{=:w} - \underbrace{\frac{1}{2}\log(2\pi\sigma^2)}_{\log Z(w)}\Big)
\end{aligned}$$

Therefore, for a scalar Gaussian distribution with known mean, we have the **sufficient statistics** $\phi(x) = -\frac{(x_i-\mu)^2}{2}$, the **natural parameters** $w = \sigma^{-2}$ and the **partition function** $Z(w) = \sqrt{\frac{2\pi}{w}}$.

(b)

We want to show that $p(w) = \mathcal{G}(w;\alpha,\beta)$ is a conjugate prior to the likelihood $p(x \mid w) = \exp\left(\phi(x)w - \log Z(w)\right)$, i.e.,

$$p(w \mid x_1,\dots,x_m) = \frac{\mathcal{G}(w;\alpha,\beta)\prod_{i=1}^m p(x_i \mid w)}{p(x_1,\dots,x_m)} = \mathcal{G}(w;\alpha_m,\beta_m).$$

We start with the numerator

$$\begin{aligned}
\mathcal{G}(w;\alpha,\beta)\prod_{i=1}^m p(x_i \mid w) &= \frac{\beta^\alpha}{\Gamma(\alpha)} w^{\alpha-1} e^{-\beta w} \prod_{i=1}^m \frac{1}{(2\pi/w)^{1/2}} e^{-\phi(x_i)w} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)(2\pi)^{m/2}}\, w^{\alpha+\frac{m}{2}-1} \exp\left(-w\left(\beta + \sum_{i=1}^m \phi(x_i)\right)\right) \quad \propto \mathcal{G}(w;\alpha_m,\beta_m)
\end{aligned}$$

with

$$\alpha_m = \alpha + \frac{m}{2} \quad \text{and} \quad \beta_m = \beta + \sum_{i=1}^m \phi(x_i) = \beta + \sum_{i=1}^m \frac{(x_i-\mu)^2}{2}.$$

Since $p(w \mid x_1,\dots,x_m)$ is a probability density on w, it has to integrate to 1 and thus

$$p(w \mid x_1,\dots,x_m) = \mathcal{G}(w;\alpha_m,\beta_m).$$

□

For completeness, we can also show that the constants work out and compute the marginal

$$\begin{aligned}
p(x_1,\dots,x_m) &= \int_0^\infty \mathrm{d}w\; \mathcal{G}(w;\alpha,\beta)\prod_{i=1}^m p(x_i \mid w) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)(2\pi)^{m/2}} \int_0^\infty \mathrm{d}w\; w^{\alpha+\frac{m}{2}-1} \exp\left(-w\left(\beta + \sum_{i=1}^m \phi(x_i)\right)\right) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)(2\pi)^{m/2}}\; \frac{\Gamma(\alpha+\frac{m}{2})}{(\beta + \sum_{i=1}^m \phi(x_i))^{\alpha+m/2}}
\end{aligned}$$

where we have deducted the integral from observing the normalization constant in the Gamma distribution. With this, the expression for the posterior becomes

$$\begin{aligned}
p(w \mid x_1,\dots,x_m) &= \frac{\cancel{\beta^\alpha}}{\cancel{\Gamma(\alpha)(2\pi)^{m/2}}}\; \frac{\cancel{\Gamma(\alpha)(2\pi)^{m/2}}}{\cancel{\beta^\alpha}}\; \frac{(\beta + \sum_{i=1}^m \phi(x_i))^{\alpha+m/2}}{\Gamma(\alpha+\frac{m}{2})} w^{\alpha+\frac{m}{2}-1} \exp\left(-w\left(\beta + \sum_{i=1}^m \phi(x_i)\right)\right) \\
&= \mathcal{G}(w;\alpha_m,\beta_m).
\end{aligned}$$

(c)

For predictions, we wish to compute

$$
\begin{aligned}
p(x_{m+1} \mid \alpha_m, \beta_m) &= \int_0^\infty \mathrm{d}w \; p(x_{m+1} \mid w)\mathcal{G}(w; \alpha_m, \beta_m) \\
&= \frac{\beta_m^{\alpha_m}}{\sqrt{2\pi}\,\Gamma(\alpha_m)} \int_0^\infty \mathrm{d}w \; w^{\alpha_m + \frac{1}{2} - 1} \exp\left(-w\left(\beta_m - \phi(x_{m+1})\right)\right) \\
&= \frac{\beta_m^{\alpha_m}}{\sqrt{2\pi}\,\Gamma(\alpha_m)} \frac{\Gamma(\alpha_m + \frac{1}{2})}{\left(\beta_m - \phi(x_{m+1})\right)^{\alpha_m + \frac{1}{2}}} \\
&= \frac{\Gamma(\alpha_m + \frac{1}{2})}{\Gamma(\alpha_m)} \frac{1}{\sqrt{2\pi\beta_m}} \left(1 + \frac{(x_{m+1} - \mu)^2}{2\beta_m}\right)^{-\alpha_m - \frac{1}{2}}
\end{aligned}
$$

$\square$

which is the so-called Student-t distribution.