Lecture Note - 06: Exponential Family, Generalize Linear Model (GLM)

Dihui Lai

March 29, 2020

Contents

1 Exponential Family

1.1 Probability Density Function

The probability density function (PDF) of an exponential family can be written in a unified format

$$f(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

There are two parameters in the function: θ and ϕ. Probability distributions like normal, Poisson or binomial can all be written in this format when we choose the right functions $a(\cdot)$, $b(\cdot)$, and $c(y, \cdot)$ and parameter θ, ϕ.

1.2 Special Case: Gaussian Distribution

For example, if we set θ as the mean and ϕ as the standard deviation:

$$\theta = \mu$$

$$\phi = \sigma$$

Make the three functions a, b, c, to be the followings:

$$a(\phi) = \phi^2 = \sigma^2$$

$$b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$$

$$c(y, \phi) = -\frac{y^2}{2\phi^2} - log(\sqrt{2\pi}\phi) = -\frac{y^2}{2\sigma^2} - log(\sqrt{2\pi}\sigma)$$

Substitute the functions into f(y), we have

$$\begin{aligned} f(y) &= \exp\left(\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right) \\ &= \exp\left(\frac{2y\mu - \frac{2\mu^2}{2} - y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right) \\ &= \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)\exp(-\log(\sqrt{2\pi}\sigma)) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

This is exactly the Gaussian distribution with mean μ and standard deviation σ

1.3 Special Case: Bernoulli Distribution

Alternatively, if we set the parameters and the functions to be the followings

$$\begin{aligned}
&\theta = \log(p) - \log(1-p) \\
&a(\phi) = 1 \\
&b(\theta) = \log(1+e^{\theta}) = -\log(1-p) \\
&c(y,\phi) = 0
\end{aligned}$$

We have the PDF function as

$$\begin{aligned}
f(y) &= \exp\left(\frac{y(\log(p) - \log(1-p)) - (-\log(1-p))}{1} + 0\right) \\
&= \exp\left(\log(p^y) + \log(1-p)^{(-y)}) + \log(1-p)^1\right) \\
&= \exp\left(\log(p^y) + \log(1-p)^{(1-y)})\right) \\
&= p^y(1-p)^{(1-y)} \\
&= \begin{cases} p, & y = 1 \\ 1-p, & y = 0 \end{cases}
\end{aligned}$$

, which is exactly the probability mass of a Bernoulli distribution.

2 Moments of Exponential Family

2.1 Mean

If a random variable Y's PDF belongs to the exponential family. We have the mean and variance of Y immediately as

$$E(Y) = \frac{d}{d\theta}b(\theta) \tag{1}$$

$$Var(Y) = a(\phi)\frac{d^2}{d\theta^2}b(\theta) \tag{2}$$

To derive the equations above, we can simply use the condition that $f(y)$ has to satisfy

$$\int_{-\infty}^{\infty} f(y)dy = 1$$

i.e.

$$\int_{-\infty}^{\infty} \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy = 1 \tag{3}$$

Take the derivative against θ on both side of the equation, we have

$$\frac{d}{d\theta}\int_{-\infty}^{\infty}\exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy = 0$$

Without proof, we assume the derivative and integration are interchangeable. We then get

$$\int_{-\infty}^{\infty}\frac{d}{d\theta}\exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy = 0$$

The the derivative against the exponential function, we have

$$\int_{-\infty}^{\infty}(y - b'(\theta))\exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy = 0$$

Rearrange the equation, we have

$$\int_{-\infty}^{\infty} y\exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy = b'(\theta)\int_{-\infty}^{\infty}\exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy$$

This is pretty nice. Because the left hand side is simply the definition of the expectation of y. On the right side, we can simplify the term using equation (3). We therefore end up having

$$E(Y) = \int_{-\infty}^{\infty} y\exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy = b'(\theta)1 = b'(\theta)$$

2.2 Variance

To get the variance of Y, we can use the same trick of taking derivative but against equation (1) or its equivalence

$$\int_{-\infty}^{\infty} y\exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy = b'(\theta)$$

i.e.

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} y\exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy = b''(\theta)$$

Again, assuming the derivative and integration is interchangeable, we have

$$\int_{-\infty}^{\infty}\frac{y(y - b'(\theta))}{a(\phi)}\exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)dy = b''(\theta)$$

We therefore have

$$E(Y^2) - (b'(\theta))^2 = a(\phi)b''(\theta)$$

or equivalently

$$Var(Y) = E(Y^2) - E(Y)^2 = a(\phi)b''(\theta)$$

3 Generalized Linear Model (GLM)

In logistic regression, the model assume the observed outcomes or values of the target variable follow binomial distribution with parameter p. The parameter p or the expectation of the distribution is modeled as a logistic function of a series of predictors $\vec{x}$ i.e. $p = \frac{1}{1+\exp(-\vec{\beta}\cdot\vec{x})}$.

3.1 Link Function

In general, our target variable could follow different types of distribution e.g. Poisson, Gaussian etc. The generalized linear model tries can be used when the target variable y follows a distribution that belongs to the exponential family.

The expectation of Y is modeled using a link function $g^{-1}(\cdot)$

$$E(Y) = b'(\theta) = g(\vec{x} \cdot \vec{\beta})$$

For simplicity, we denote $\eta = \vec{x} \cdot \vec{\beta}$ and thus

$$E(Y) = b'(\theta) = g(\eta) \tag{4}$$

If we want to estimate the parameter θ, we can solve the equation (4) by taking the inverse of the function $b'(\cdot)$. If we choose the inverse link function $g(\cdot)$ to be the same as $b'(\cdot)$

$$g(\cdot) = b'(\cdot) \tag{5}$$

we get a model with nice property where $\eta = \theta$. A link function chosen this way is called a canonical link function.

3.2 Canonical Link Function

Bernoulli Distribution Use Bernoulli distribution as an example, we have $b'(\theta) = \frac{e^\theta}{1+e^\theta}$. The inverse canonical link function is

$$g(\eta) = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}}$$

As we known,

$$p = b'(\theta) = \frac{1}{1 + e^{-\eta}}$$

The equation looks a lot like the probability model in the logistic regression model. In fact, as we will learn in the following section, GLM in the special case of a Bernoulli distribution is equivalent to a logistic regression.

Gaussian Distribution: in the case of a Gaussian distribution, $b'(\theta) = \theta$, which is an identity function. Choose the link function as the identify function, we have

$$g(\eta) = \eta$$

Since we have $\mu = \theta$, the canonical link function of a Gaussian distribution models the expectation of the target variable as of the following

$$\mu = \eta = \vec{x} \cdot \vec{\beta}$$

This looks a lot like the linear regression model, except that we don not know how GLM estimates the error function yet. In the following sections, we are going to discuss how to use MLE to optimize the a GLM.

3.3 Maximum Likelihood Estimation

3.3.1 Likelihood Function

A good model represent a dataset as close as it can. We can uses maximum likelihood estimation to optimize the $\vec{\beta}$ in GLM.

Considering one data entry $(\vec{x}^i, y^i)$, the corresponding log-likelihood function is

$$\ell^i = \frac{y^i\theta^i - b(\theta^i)}{a(\phi)} + c^i(y^i, \phi)$$

Since only θ^i is dependent on $\vec{\beta}$, to maximize the function is equivalent to minimize the function $\ell^i = -2\left[y^i\theta^i - b(\theta^i)\right]$. The log-likelihood for the whole data set is a summation of each individual ℓ^i. Overall, we need to minimize the following log-likelihood.

$$\ell = -2\sum_{i=1}^{N}\left[y^i\theta^i - b(\theta^i)\right] \tag{6}$$

In the case of Bernoulli distribution we have $\ell = y^i\theta^i - \log\left(1 + \exp(\theta^i)\right)$

3.3.2 Gradient

To maximize the likelihood ℓ, we need to have

$$\frac{\partial \ell}{\partial \beta^j} = 0$$

By using equation (6), we end up having

$$\frac{\partial \ell}{\partial \beta^j} = -2 \sum_{i=1}^{N} \frac{\partial}{\partial \beta^j} \left[y^i \theta^i - b(\theta^i) \right] \tag{7}$$

$$= -2 \sum_{i=1}^{N} \left[y^i - b'(\theta^i) \right] \frac{\partial \theta^i}{\partial \beta^j} = 0 \tag{8}$$

If we use canonical link function, we know $\eta = \theta$, the equation can be simplified as below

$$\frac{\partial \ell}{\partial \beta^j} = -2 \sum_{i=1}^{N} \left[y^i - b'(\theta^i) \right] \frac{\partial \eta^i}{\partial \beta^j} \tag{9}$$

$$= -2 \sum_{i=1}^{N} \left[y^i - b'(\theta^i) \right] x_j^i = 0 \tag{10}$$

In the case of a Bernoulli distribution, we have $b'(\theta) = p$

$$\frac{\partial \ell}{\partial \beta^j} = \sum_{i=1}^{N} [y^i x_j^i - p^i x_j^i]$$

This is equivalent to the corresponding equation using MLE in logistic regression. In another word, GLM is equivalent to logistic regression when we assume Bernoulli/binomial distribution of the target variable and use canonical link function.

In general, $\eta \neq \theta$ and we will not have an estimator as simple as equation (10). We need to solve equation (8) instead. In order to do this numerically, we need to figure out the value of $\frac{\partial \theta^i}{\partial \beta^j}$ using βs and xs, or η equivalently.

In order to get the relationship between θ and β, we start from the equation (4), where the relationship of the two variables are defined by GLM. By taking the derivative against βs, we have

$$\frac{\partial}{\partial \beta_j} b'(\theta) = g'(\eta) \frac{\partial \eta}{\partial \beta} \tag{11}$$

$$b''(\theta) \frac{\partial \theta}{\partial \beta} = g'(\eta) \frac{\partial \eta}{\partial \beta_j} \tag{12}$$

Therefore we get

$$\frac{\partial \theta}{\partial \beta} = \frac{g'(\eta)}{b''(\theta)} \frac{\partial \eta}{\partial \beta_j} \tag{13}$$

Inserting this back to equation (8), we have

$$\frac{\partial \ell}{\partial \beta_j} = -2 \sum_{i=1}^{N} \left[y^i - b'(\theta^i) \right] \frac{g'(\eta^i)}{b''(\theta^i)} x_j^i = 0 \tag{14}$$

Notice that we have an extra weight term $\frac{g'(\eta)}{b''(\theta)}$ comparing to equation (10). To simplify the notation, we denote $b'(\theta) = \mu$ and $b''(\theta) = V(\mu)$. We therefore have

$$\frac{\partial \ell}{\partial \beta_j} = -2 \sum_{i=1}^{N} (y^i - \mu^i) \frac{g'(\eta^i)}{V(\mu^i)} x_j^i = 0 \tag{15}$$

3.3.3 Hessian

We can use Newton-Raphson method to solve the MLE problem formulated above, however, we need to calculate the Hessian first.

The Hessian matrix can be estimated by taking the derivative of the gradient, noting here μ is a function of θ and therefore dependent on β.

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j} = -2 \sum_{i=1}^{N} \frac{\partial}{\partial \beta_k} \left[(y^i - \mu^i) \frac{g'(\eta^i)}{V(\mu^i)} x_j^i \right] \tag{16}$$

$$= -2 \sum_{i=1}^{N} \left[-\frac{\partial \mu^i}{\partial \beta_k} \frac{g'(\eta^i)}{V(\mu^i)} + (y^i - \mu^i) \frac{g''(\eta^i) V(\mu^i) \frac{\partial \eta^i}{\partial \beta_k} - g'(\eta^i) V'(\mu^i) \frac{\partial \mu^i}{\partial \beta_k}}{V(\mu^i)^2} \right] x_j^i \tag{17}$$

$$= 2 \sum_{i=1}^{N} \left[g'(\eta^i) \frac{g'(\eta^i)}{V(\mu^i)} - (y^i - \mu^i) \frac{g''(\eta^i) V(\mu^i) - g'(\eta^i)^2 V'(\mu^i)}{V(\mu^i)^2} \right] \frac{\partial \eta^i}{\partial \beta_k} x_j^i \tag{18}$$

$$= 2 \sum_{i=1}^{N} \left[g'(\eta^i) \frac{g'(\eta^i)}{V(\mu^i)} - (y^i - \mu^i) \frac{g''(\eta^i) V(\mu^i) - g'(\eta^i)^2 V'(\mu^i)}{V(\mu^i)^2} \right] x_k^i x_j^i \tag{19}$$

Going from equation (17) to (18), we have used the fact that $\frac{\partial \mu}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} b'(\theta) = g'(\eta) \frac{\partial \eta}{\partial \beta}$