Predictive Analytics
Regression and Classification
Lecture 10 : Part 1

Sourish Das

Chennai Mathematical Institute

Aug-Nov, 2019

Natural Exponential Family

- Suppose $y_1, y_2, \cdots, y_n$ are independent observations where y_i has desnsity from natural exponential family

$$f(y_i|\theta_i) = h(y_i)\exp\{(\eta(\theta_i)T(y_i) - \psi(\theta_i))\},$$

where $i = 1, 2, \cdots, n$.

- $\eta(\theta_i)$ is known as canonical parameter

- $\psi(.)$ and $h(.)$ are known function

Binomial distribution

- Suppose $y_1, y_2, \cdots, y_n \sim Bin(m, \theta_i)$

$$\begin{aligned} f(y_i|\theta_i) &= {}^mC_{y_i}\theta_i^{y_i}(1-\theta_i)^{m-y_i}, \\ &= {}^mC_{y_i}\left(\frac{\theta_i}{1-\theta_i}\right)^{y_i}(1-\theta_i)^m \\ &= \underbrace{{}^mC_{y_i}}_{h(y_i)}\exp\left\{\underbrace{\log\left(\frac{\theta_i}{1-\theta_i}\right)}_{\eta(\theta_i)}y_i - \underbrace{m\log(1-\theta_i)}_{\psi(\theta_i)}\right\} \end{aligned}$$

where $i = 1, 2, \cdots, n$.

- $h(y_i) =^m C_{y_i}$
- $\eta(\theta_i) = \log(\frac{\theta_i}{1-\theta_i})$
- $T(y_i) = y_i$
- $\psi(\theta_i) = -m\log(1-\theta_i)$

Poisson distribution

- Suppose $y_1, y_2, \cdots, y_n \sim Poisson(\theta_i)$

$$f(y_i|\theta_i) = \frac{\theta_i^{y_i}}{y_i!}\exp\{-\theta_i\}, = \frac{1}{y_i!}\exp\{\log(\theta_i)y_i - \theta_i\}$$

where $i = 1, 2, \cdots, n$.

  - $h(y_i) = \frac{1}{y_i!}$
  - $\eta(\theta_i) = \log(\theta_i)$
  - $T(y_i) = y_i$
  - $\psi(\theta_i) = \theta_i$

Normal distribution

- Suppose $y_1, y_2, \cdots, y_n \sim Normal(\theta_i, \sigma^2)$ (σ^2 is known)

$$\begin{aligned} f(y_i|\theta_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\theta_i)^2}, \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\{-\frac{y_i^2}{2\sigma^2}\} \times \exp\{\theta_i y_i - \frac{\theta_i^2}{2}\} \end{aligned}$$

where $i = 1, 2, \cdots, n$.

  - $h(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\{-\frac{y_i^2}{2\sigma^2}\}$
  - $\eta(\theta_i) = \theta_i$
  - $T(y_i) = y_i$
  - $\psi(\theta_i) = \frac{\theta_i^2}{2\sigma^2}$

cmi

Generalized Linear Model

1. **Random Component** $y_i \sim NEF(\theta_i)$ with pdf

$$f(y_i|\theta_i) = h(y_i)\exp\{(\eta(\theta_i)T(y_i) - \psi(\theta_i))\},$$

   where $i = 1, 2, \cdots, n$.

2. **Link function**: $\eta(\theta_i) = z_i$

3. **Systematic component**: $z_i = \mathbf{x}_i^T \boldsymbol{\beta}$

Generalized Linear Model (GLM)

1. **Random Component** $y_i \sim NEF(\theta_i)$ with pdf

$$f(y_i|\theta_i) = h(y_i)\exp\{(\eta(\theta_i)T(y_i) - \psi(\theta_i))\},$$

   where $i = 1, 2, \cdots, n$.

2. **Systematic component**: $\eta(\theta_i) = \mathbf{x}_i^T \boldsymbol{\beta}$

Regression with GLM

1. **Random Component** $y_i \sim N(\theta_i, \sigma^2)$ with pdf

$$
\begin{aligned}
f(y_i|\theta_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\theta_i)^2}, \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\{-\frac{y_i^2}{2\sigma^2}\} \times \exp\{\theta_i y_i - \frac{\theta_i^2}{2}\}
\end{aligned}
$$

where $i = 1, 2, \cdots, n$.

2. **Systematic component**: $\eta(\theta_i) = \theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$

Count Regression with GLM

1. **Random Component** $y_i \sim Poisson(\theta_i)$ with pf

$$\begin{aligned} f(y_i|\theta_i) &= \frac{\theta_i^{y_i}}{y_i!} \exp\{-\theta_i\}, \\ &= \frac{1}{y_i!} \exp\{\log(\theta_i) y_i - \theta_i\} \end{aligned}$$

   where $i = 1, 2, \cdots, n$.

2. **Systematic component**: $\eta(\theta_i) = \log(\theta_i) = \mathbf{x}_i^T \boldsymbol{\beta}$

Calissification with GLM

1. **Random Component** $y_i \sim Bin(1, \theta_i)$ with pdf

$$
\begin{aligned}
f(y_i|\theta_i) &= \theta_i^{y_i}(1-\theta_i)^{1-y_i}, \\
&= \exp\left\{\log\left(\frac{\theta_i}{1-\theta_i}\right)y_i - \log(1-\theta_i)\right\}
\end{aligned}
$$

where $i = 1, 2, \cdots, n$.

2. **Systematic component**: $\eta(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta}$

Likelihood function of GLM

- Negative log-Likelihood function of GLM

$$
\begin{aligned}
-\log L &= -\sum_{i=1}^{n} \log(f(y_i|\theta_i)) \\
&= -\sum_{i=1}^{n} \log(f(y_i|\eta^{-1}(\mathbf{x}_i^T \beta)))
\end{aligned}
$$

- MLE of β of GLM

$$
\hat{\beta}_{MLE} = \text{argmin}_{\beta} \left[-\sum_{i=1}^{n} \log(f(y_i|\eta^{-1}(\mathbf{x}_i^T \beta))) \right]
$$

Implement GLM with R

- **Regression**:

```
> stats::glm(y~x1+x2
+             ,family=gaussian(link = "identity")
+             ,data=data_nm)
```

- **clssification with logistic regression**:

```
> stats::glm(y~x1+x2
+             ,family=binomial(link = "logit")
+             ,data=data_nm)
```

- **count / Poisson regression**:

```
> stats::glm(y~x1+x2
+             ,family=poisson(link = "log")
+             ,data=data_nm)
```

Thank You

sourish@cmi.ac.in