清華大學
Tsinghua University

Recitation : Two examples of GLM

Yuanbo Tang, Yang Li*
Tsinghua-Berkeley Shenzhen Institute, Tsinghua University

Generalized Linear Models (GLM)

Generalized Linear Models (GLM) are a class of models in statistics designed to *address the limitations of linear regression models :*

- Non-Normally Distributed Target Variables:
  GLM allows the target variable to *follow an exponential family distribution, not just a normal distribution*. This enables GLM to handle binary data (e.g., logistic regression) , count data (e.g., Poisson regression)…

- Nonlinear Relationships:
  By introducing a link function, GLM can capture *nonlinear relationships* between variables. For example, logistic regression uses a logit link function to address nonlinear relationships in binary classification problems.

Generalized Linear Models (GLM)

Generalized Linear Models (GLM) are a class of models in statistics designed to *address the limitations of linear regression models.*

Formal GLM assumptions & design decisions:

1. $y|x;\theta \sim \text{ExponentialFamily}(\eta)$
   e.g. Gaussian, Poisson, Bernoulli, Multinomial, Beta ...
2. The hypothesis function $h(x)$ is $\mathbb{E}[T(y)|x]$
   e.g. When $T(y) = y$, $h(x) = \mathbb{E}[y|x]$
3. The natural parameter η and the inputs x are related linearly:

   η is a number:

   $$\eta = \theta^T x$$

   η is a vector:

   $$\eta_i = \theta_i^T x \quad \forall i = 1, \ldots, n \quad \text{or} \quad \eta = \Theta^T x$$

Generalized Linear Models (GLM)

Generalized Linear Models (GLM) are a class of models in statistics designed to *address the limitations of linear regression models.*

Relate natural parameter η to distribution mean $\mathbb{E}[T(y)|x]$:

- **Canonical response function** g gives the mean of the distribution

$$g(\eta) = \mathbb{E}[T(y)|x]$$

  a.k.a. the "mean function"

- g^{-1} is called the **canonical link function**

$$\eta = g^{-1}(\mathbb{E}[T(y)|x])$$

Review: Exponential Family

The Exponential Family is an important class of probability distributions in statistics. They have a *unified mathematical form* and are widely used in statistical modeling and machine learning.

A class of distributions is in the **exponential family** if its density can be written in the *canonical form*:

$$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)}$$

- y: random variable
- η : natural/canonical parameter (that depends on distribution parameter(s))
- $T(y)$: sufficient statistic of the distribution
- $b(y)$: a function of y
- $a(\eta)$: log partition function (or "cumulant function")

Review: Exponential Family

A class of distributions is in the **exponential family** if its density can be written in the *canonical form*:

$$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)}$$

Bernoulli Distribution

Bernoulli(ϕ): a distribution over $y \in \{0, 1\}$, such that

$$p(y;\phi) = \phi^y(1-\phi)^{1-y}$$

- $\eta = \log\left(\frac{\phi}{1-\phi}\right)$
- $b(y) = 1$
- $T(y) = y$
- $a(\eta) = \log(1 + e^{\eta})$

Review: Exponential Family

A class of distributions is in the **exponential family** if its density can be written in the *canonical form*:

$$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)}$$

1. **Original PMF**:

$$p(y;\phi) = \phi^y (1-\phi)^{1-y}$$

2. **Taking the logarithm**:

$$\log p(y;\phi) = y\log(\phi) + (1-y)\log(1-\phi)$$

3. **Rewriting the logarithm**:

$$\log p(y;\phi) = y\log\left(\frac{\phi}{1-\phi}\right) + \log(1-\phi)$$

4. **Exponentiating to get back to the original form**:

$$p(y;\phi) = \exp\left(y\log\left(\frac{\phi}{1-\phi}\right) + \log(1-\phi)\right)$$

- $\eta = \log\left(\frac{\phi}{1-\phi}\right)$
- $b(y) = 1$
- $T(y) = y$
- $a(\eta) = \log(1+e^{\eta})$

Review: Exponential Family

A class of distributions is in the **exponential family** if its density can be written in the *canonical form*:

$$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)}$$

Gaussian Distribution (unit variance)

Probability density of a Gaussian distribution $\mathcal{N}(\mu, 1)$ over $y \in \mathbb{R}$:

$$p(y;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2}\right)$$

- $\eta = \mu$
- $b(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$
- $T(y) = y$
- $a(\eta) = \frac{1}{2}\eta^2$

Review: Exponential Family

A class of distributions is in the **exponential family** if its density can be written in the *canonical form*:

$$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)}$$

1. **Original PDF**:

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

2. **Expanding the Quadratic Term**:

$$-\frac{(y-\mu)^2}{2} = -\frac{y^2}{2} + \mu y - \frac{\mu^2}{2}$$

3. **Rewriting the PDF**:

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} + \mu y - \frac{\mu^2}{2}\right)$$

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

- $\eta = \mu$
- $b(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$
- $T(y) = y$
- $a(\eta) = \frac{1}{2}\eta^2$

GLM example: ordinary least square

Apply GLM construction rules:

1. Let $y|x;\theta \sim N(\mu, 1)$

$$\eta = \mu, \ T(y) = y$$

2. Derive hypothesis function:

$$\begin{aligned} h_\theta(x) &= \mathbb{E}\left[T(y)|x;\theta\right] \\ &= \mathbb{E}\left[y|x;\theta\right] \\ &= \mu = \eta \end{aligned}$$

3. Adopt linear model $\eta = \theta^T x$:

$$h_\theta(x) = \eta = \theta^T x$$

Canonical response function: $\mu = g(\eta) = \eta$ (identity)
Canonical link function: $\eta = g^{-1}(\mu) = \mu$ (identity)

GLM example 2: logistic regression

Apply GLM construction rules:

1. Let $y|x;\theta \sim \text{Bernoulli}(\phi)$

$$\eta = \log\left(\frac{\phi}{1-\phi}\right),\ T(y) = y$$

2. Derive hypothesis function:

$$\begin{aligned} h_\theta(x) &= \mathbb{E}\left[T(y)|x;\theta\right] \\ &= \mathbb{E}\left[y|x;\theta\right] \\ &= \phi = \frac{1}{1+e^{-\eta}} \end{aligned}$$

3. Adopt linear model $\eta = \theta^T x$:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

Canonical response function: $\phi = g(\eta) = \text{sigmoid}(\eta)$
Canonical link function : $\eta = g^{-1}(\phi) = \text{logit}(\phi)$

Conclusion

In summary

- GLMs help us extend linear regression by allowing for different types of data and relationships.
- Understanding GLMs and the Exponential Family is important for choosing the right models in data analysis.

Thanks

Thank you for listening!

If you have any question,
please come and discuss with me for more details.