Generalized Linear Models

Baboo J. Cui

July 30, 2019

We have seen regression and classification examples in which Gaussian and Bernoulli distribution are involved. All of the cases can be expanded to generalized linear models(GLMs).

1 The Exponential Family

Exponential family distribution can be generally written as

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

- natural parameter(canonical parameter): η
- sufficient statistics: $T(y)$, often $T(y) = y$
- log partition function: $a(\eta)$, related to normalization
- normalization constant: $e^{-a(\eta)}$, make the integrate over y to be 1
- a family of distribution is determined by T, a and b, and η is the parameter
- any term for $b(y)$?

Let's take Bernoulli distribution as a example(easy). Suppose $y \sim \mathcal{B}(\phi)$, the distribution can be written as

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\ln\phi + (1-y)\ln(1-\phi)) \\ &= \exp\left(\left(\ln\frac{\phi}{1-\phi}\right)y + \ln(1-\phi)\right) \end{aligned}$$

Clearly,

$$\begin{aligned} \eta &= \ln\frac{\phi}{1-\phi} \qquad \Longrightarrow \qquad \phi = \frac{1}{1+e^{-\eta}} \\ T(y) &= y \\ a(\eta) &= -\ln(1-\phi) = \ln(1+e^{\eta}) \\ b(y) &= 1 \end{aligned}$$

The following distributions are members of the exponential family: multinomial, Poisson, gamma, exponential, beta and Dirichlet distributions.

2 Constructing GLMs

This part discuss how to construct GLM if the distribution is in exponential family, like using Poisson distribution to estimate the number of visitor in a store. To derive the model, 3 assumptions are made:

- $y|x;\theta \sim \text{ExpFamily}(\eta)$
- in most case $T(y) = y$, we want $h(x) = E[y|x]$
- natural parameter η is linearly related to input x, that is $\eta = \theta^T x$

This will allow us to derive a very elegant class of learning algorithm. And the target variable in GLM is called response variable.

2.1 Ordinary Least Square

Let's model the conditional distribution of y given x as Gaussian $\mathcal{N}(\mu, \sigma^2)$. We have

$$\begin{aligned}
h_\theta(x) &= E[y|x;\theta] && \text{recall we want: } h(x) = E[y|x] \\
&= \mu && \text{property of normal distribution} \\
&= g(\eta) = \eta && \text{from formulation of GLM(identify function)} \\
&= \theta^T x && \eta \text{ is a linear function of } \theta
\end{aligned}$$

VIP: h can be written as a function of η which is linearly related to θ.

2.2 Logistic Regression

Here we talk about binary classification, so $y \in \{0, 1\}$. Bernoulli distribution should be chosen so that

$$y|x;\theta \sim \mathcal{B}(\phi)$$

and expectation can be written as

$$E[y|x;\theta] = \phi$$

Follow what has done for least square case, we have

$$\begin{aligned}
h_\theta(x) &= E[y|x;\theta] && \text{recall we want: } h(x) = E[y|x] \\
&= \phi && \text{property of Bernoulli distribution} \\
&= g(\eta) = \frac{1}{1+e^{-\eta}} && \text{GLM on Bernoulli(logistic function)} \\
&= \frac{1}{1+e^{-\theta^T x}} && \eta \text{ is a linear function of } \theta
\end{aligned}$$

Note that GLM can directly give the hypothesis function. And more:

- canonical response function: function g, where $g(\eta) = E[T(y);\eta]$
- canonical link function: g^{-1}

3 Softmax Regression

Consider a classification problem in which the response variable can take on any one of k values, so $y \in \{1, 2, \cdots, k\}$. It will follow multinomial distribution.
To parametrize multinomial distribution over k possible outcomes, $k-1$ independent parameters are required. So we have:

$$\begin{aligned} P(y = i; \phi) &= \phi_i \quad \text{where } i \neq k \\ P(y = k; \phi) &= 1 - \sum_{i=1}^{k-1} \phi_i = \phi_k \end{aligned}$$

Define $T(y) \in \mathbb{R}^{k-1}$, which is not scalar anymore, as

$$(T(y))_i = \begin{cases} 1 & \text{if } y = i \\ 0 & \text{otherwise} \end{cases}$$

where subscript i represent the i-th element in the matrix. And $T(k)$ is a zero matrix. And indicator function $1\{\cdot\}$ is defined as

$$\begin{aligned} 1\{\text{True}\} &= 1 \\ 1\{\text{False}\} &= 0 \end{aligned}$$

So $T(y)$ can also be denote as

$$(T(y))_i = 1\{y = i\}$$

And further

$$E[(T(y)_i)] = P(y = i) = \phi_i$$

To derive multinomial probability distribution is a member of exponential family:

$$\begin{aligned} p(y; \phi) &= \phi_1^{1\{y=1\}} \phi_2^{1\{y=2\}} \cdots \phi_k^{1\{y=k\}} \\ &= b(y) \exp(\eta^T T(y) - a(\eta)) \end{aligned}$$

where

$$\begin{aligned} \eta &= \begin{bmatrix} \ln(\phi_1/\phi_k) \\ \ln(\phi_2/\phi_k) \\ \vdots \\ \ln(\phi_{k-1}/\phi_k) \end{bmatrix} \\ a(\eta) &= -\ln(\phi_k) \\ b(y) &= 1 \end{aligned}$$

The link function is given by

$$\eta_i = \eta(\phi) = \ln \frac{\phi_i}{\phi_k}$$

The response function is

$$\phi_i = \phi(\eta) = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$$

And it is called softmax function. Take previous assumption that $\eta = \theta^T x$, we have

$$\begin{aligned}
p(y = i|x;\theta) &= \phi_i \\
&= \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}} \\
&= \frac{e^{\theta_i^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}}
\end{aligned}$$

This is called softmax regression, it is generalization of logistic regression, and the hypothesis will output the estimated probability

$$\begin{aligned}
h_\theta(x) &= E[T(y)|x;\theta] \\
&= \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{bmatrix} \\
&= \begin{bmatrix} \frac{e^{\theta_1^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \\ \frac{e^{\theta_2^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \\ \vdots \\ \frac{e^{\theta_{k-1}^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \end{bmatrix}
\end{aligned}$$

This will output the estimated probability of $P(y|x;\theta)$. To do the parameter fitting, the likelihood function is(for m training example)

$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)$$

It can be maximized by maximizing the corresponding log function as what we did earlier.

4 Summary

The GLM algorithm can be summarized as the following:

1. express $h_\theta(x) = E[y|x;\theta]$ as the purpose
2. write the expectation as a function of distribution parameter

3. according to exponential family, write expectation as natural parameter η
4. model h_θ now can be written as function of x since $\eta = \theta^T x$
5. use whatever the way to maximize the expectation in terms of θ
6. when θ is optimized, the training is finished

Recall:

- link function express natural parameters in terms of distribution parameters
- response function expresses distribution parameters in terms of natural parameters