Generalized Linear Model (GLM)

June 8, 2023

1 Introduction

Generalized linear models are an abstraction of various types of probablistic models used for classification and regression problems.
GLMs assume $Y|X;\theta \sim$ exponential family (see background). The goal of a GLM is to predict y from x, therefore, it is a supervised learning model. Various probablistic distributions can be represented as an exponetial family and thus can be approximated via a GLM.
Here is a table of some common exponential family distributions and their use cases as a GLM:

Type	Distribution	Use Case
Linear Regression	Gaussian	Real Valued Data
Logisitic Regression	Bernoulli	Binary classification
Softmax	Multinomial	Multiclass classification
Poisson Regression	Poisson	Natural Number
Exponential Regression	Exponential	Real Positive Number
Gamma Regression	Gamma	Real Positive Number
Beta Regression	Beta	Probabilty Distribution
Dirichlet Regression	Dirichlet	Probabilty Distribution

2 Background

Exponential Families are distributions whose pdf can be written in the form:

$$p(y|\eta) = b(y)e^{\eta^t T(y) - A(\eta)}$$

$$y : \text{target variable}$$

$$\eta : \text{natural parameter}$$

$$T(y) : \text{sufficient statistic}$$

$$b(y) : \text{base measure}$$

$$A(\eta) : \text{log partition function}$$

Exponential families have the interesting property, $E[T(y)|\eta] = \frac{\partial A}{\partial \eta}$, that is quite useful for GLMs.
Proof that $E[T(y)|\eta] = \frac{\partial A}{\partial \eta}$:

$$E[T(y)|\eta] = \int T(y)\dot{p}(y|\eta)dy$$

$$\text{for simplicity } \eta \in \mathbb{R}$$

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

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

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

$$\frac{\partial p(y;\eta)}{\partial \eta} = p(y;\eta)(T(y) - \frac{\partial A(\eta)}{\partial \eta})$$

$$\frac{\partial}{\partial \eta}\int p(y;\eta)dy = \int \frac{\partial p(y;\eta)}{\partial \eta}dy$$

$$\frac{\partial}{\partial \eta}\int p(y;\eta)dy = \int p(y;\eta)(T(y) - \frac{\partial A(\eta)}{\partial \eta})dy$$

$$\frac{\partial}{\partial \eta}\int p(y;\eta)dy = \int p(y;\eta)T(y)dy - \int p(y;\eta)\frac{\partial A(\eta)}{\partial \eta}dy$$

$$\text{Because } p(y|\eta) \text{ is a pdf } \int p(y|\eta) = 1$$

$$\frac{\partial}{\partial \eta}\int p(y;\eta)dy + \int p(y;\eta)\frac{\partial A(\eta)}{\partial \eta}dy = \int p(y;\eta)T(y)dy$$

$$\frac{\partial}{\partial \eta}1 + \frac{\partial A(\eta)}{\partial \eta}\int p(y;\eta)dy = \int p(y;\eta)T(y)dy$$

$$0 + \frac{\partial A(\eta)}{\partial \eta} = \int p(y;\eta)T(y)dy$$

$$\frac{\partial A(\eta)}{\partial \eta} = E[T(y);\eta]$$

Here are some examples of common distributions written as a GLM:

2.1 Gaussian

$$p(y;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(y-\mu)^2}{2\sigma^2}}$$

$$\text{for simplicity } \sigma^2 = 1$$

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

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

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

$$\text{we can clearly see that this is a GLM}$$

$$A(\eta) = \frac{\mu^2}{2}$$

$$b(y) = \frac{1}{\sqrt{2\pi}}e^{\frac{-y^2}{2}}$$

$$T(y) = y$$

$$\eta = \mu$$

$$\frac{\partial A\eta}{\partial \eta} = \eta$$

2.2 Bernoulli

$$p(y;\phi) = \phi^y(1-\phi)^{1-y}; y \in \{0,1\}$$
$$= e^{yln(\phi)+(1-y)ln(1-\phi)}$$
$$= e^{yln(\frac{\phi}{1-\phi})+ln(1-\phi)}$$

we can clearly see that this is a GLM

$$A(\eta) = -ln(1-\phi)$$
$$b(y) = 1$$
$$T(y) = y$$
$$\eta = ln(\frac{\phi}{1-\phi})$$
$$e^\eta = \frac{\phi}{1-\phi}$$
$$e^\eta - e^\eta\phi = \phi$$
$$e^\eta = \phi + e^\eta\phi$$
$$\frac{e^\eta}{1+e^\eta} = \phi$$
$$\frac{e^\eta}{1+e^\eta} \cdot \frac{e^-n}{e}^{-n} = \phi$$
$$\frac{1}{1+e^{-\eta}} = \phi$$

This is the sigmoid function! $\sigma(x) = \dfrac{1}{1+e^{-x}}$

$$\frac{\partial A\eta}{\partial \eta} = \frac{\partial}{\partial \eta} - ln(1-\sigma(\eta))$$
$$\frac{\partial A\eta}{\partial \eta} = \frac{1}{1-\sigma(\eta)}\frac{\partial}{\partial \eta}(1+\sigma(\eta))$$
$$\frac{\partial A\eta}{\partial \eta} = \frac{1}{1-\sigma(\eta)}(\sigma(\eta)(1-\sigma(\eta)))$$
$$\frac{\partial A\eta}{\partial \eta} = \sigma(\eta) = \phi$$

2.3 Multinomial

Note that Bernoulli is just a special case of Multinomial

$$\phi \in \mathbb{R}^{k-1}, y \in 1, 2..., k$$

let ϕ_i denote the ith entry of ϕ

Note: ϕ_k is dependent on $\phi_{1...k-1}$ so: $\phi_k = (1 - \sum_{i=1}^{k-1} \phi_i)$

$$p(y = i; \phi) = \phi_i$$
$$p(y;\phi) = \prod_{i=1}^{k} \phi_i^{\mathbb{1}\{y=i\}}$$

Where $\mathbb{1}\{y = i\}$ is the indicator function

$$p(y;\phi) = e^{ln(\prod_{i=1}^{k} \phi_i^{\mathbb{1}\{y=i\}})}$$

$$p(y;\phi) = e^{\sum_{i=1}^{k} \mathbb{1}\{y=i\}ln(\phi_i)}$$

$$p(y;\phi) = e^{\sum_{i=1}^{k-1} \mathbb{1}\{y=i\}ln(\phi_i)+\mathbb{1}\{y=k\}ln(\phi_k)}$$

$$p(y;\phi) = e^{\sum_{i=1}^{k-1} \mathbb{1}\{y=i\}ln(\phi_i)+(1-\sum_{i=1}^{k-1} \mathbb{1}\{y=i\})ln(\phi_k)}$$

$$p(y;\phi) = e^{\sum_{i=1}^{k-1} \mathbb{1}\{y=i\}ln(\phi_i)+ln(\phi_k)-\sum_{i=1}^{k-1} \mathbb{1}\{y=i\}ln(\phi_k)}$$

$$p(y;\phi) = e^{\sum_{i=1}^{k-1} \mathbb{1}\{y=i\}ln(\frac{\phi_i}{\phi_k})+ln(\phi_k)}$$

we can clearly see that this is a GLM

$$A(\eta) = -ln(\phi_k)$$

$$b(y) = 1$$

$$T(y)_i = \mathbb{1}\{y=i\}; T : \mathbb{R} \to \mathbb{R}^{k-1}$$

$$\eta_i = ln(\frac{\phi_i}{\phi_k}); \eta \in \mathbb{R}^{k-1}$$

$$e^{\eta_i} = \frac{\phi_i}{\phi_k}$$

$$\phi_k e^{\eta_i} = \phi_i$$

$$\phi_k \sum_{i=1}^{k} e^{\eta_i} = \sum_{i=1}^{k} \phi_i = 1$$

$$\phi_k = \frac{1}{\sum_{i=1}^{k} e^{\eta_i}}$$

$$\frac{e^{\eta_i}}{\sum_{i=1}^{k} e^{\eta_i}} = \phi_i$$

$$\frac{\partial A\eta_i}{\partial \eta_i} = \frac{\partial}{\partial \eta_i} - ln(\phi_k)$$

$$\frac{\partial A\eta_i}{\partial \eta_i} = \frac{\partial}{\partial \eta_i} - ln(\frac{1}{\sum_{i=1}^{k} e^{\eta_i}})$$

$$\frac{\partial A\eta_i}{\partial \eta_i} = \frac{\partial}{\partial \eta_i} ln(\sum_{i=1}^{k} e^{\eta_i})$$

$$\frac{\partial A\eta_i}{\partial \eta_i} = \frac{1}{\sum_{i=1}^{k} e^{\eta_i}} \frac{\partial}{\partial \eta_i} \sum_{i=1}^{k} e^{\eta_i}$$

$$\frac{\partial A\eta_i}{\partial \eta_i} = \frac{e^{\eta_i}}{\sum_{i=1}^{k} e^{\eta_i}} = \phi_i$$

3 How it works

The model works by estimating the relationship between y and x by modifying the parameter θ. We assume the following:

- $p(y|x;\theta) \sim$ Exponential Family
- $\eta = \theta^t x$
- Each example is independently and identically distributed

We then use gradient ascent to maximize the log likelihood of θ on the distribtion. The GLM makes the prediction by using the expected value of y given our x. i.e. our hypothesis function is equal to E[T(y) — x; θ]. Conveniently, as proven above, E[T(y) — x; θ] = $\frac{\partial A(\eta)}{\partial \eta}$ = h(x) of the given exponential family. With h(x) being shorthand for our hypothesis function. Lets go back and look at common examples of our hypothesis function.

3.1 Gaussian

$$x, \theta \in R^{n+1}$$
$$x_1 = 1, \text{ for the bias term}$$
$$h(x) = \eta = \theta^t x$$

Where n is the number of features.

3.2 Bernoulli

$$x, \theta \in R^{n+1}$$
$$x_1 = 1, \text{ for the bias term}$$
$$h(x) = \sigma(\eta) = \sigma(\theta^t x)$$

3.3 Softmax

$$x \in R^{n+1}, \theta \in R^{(k-1)x(n+1)}$$
$$\frac{\partial A \eta_i}{\partial \eta_i} = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}} = \frac{e^{\theta_i^t x}}{\sum_{j=1}^{k} e^{\theta_j^t} x}$$

4 Model Space

The model space is just the set of all possible weights θ. Theta is usually given by $\theta \in \mathbb{R}^{n+1}$.

5 Score Function

We score the model based upon the likelihood of θ. Usually we use log-likelihood as it is easier to compute. Since the data is IID:

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

Where m is the number of examles.

6 Search Method

Generally, we peform maximum likelihood estimation on θ to search over the model space. GLMs have another interseting property that makes search easy. For all GLMs the following is true:

$$\frac{\partial ln(\mathcal{L}(\theta))}{\partial \theta_i} = \sum_{j=1}^{m} (T(y^j) - h(x^j))x_i^j$$

Therefore, using gradient ascent we get, $\theta_j = \theta_{j-1} + \alpha \sum_{j=1}^{n} (T(y^j) - h(x^j))x_i^j$.

Note: this is the common update rule for least squares ($\theta_j = \theta_{j-1} + \alpha \sum_{j=1}^{n} (y - \theta^t x)x_i^j$)
Also for simplicity $\alpha = \frac{\beta}{m}$. Where $\beta \in \mathbb{R}$. α is also known as the learning rate and controls how much theta changes with each iteration of gradient ascent.

I will prove $\frac{\partial ln(\mathcal{L}(\theta))}{\partial \theta_i} = \sum_{j=1}^{m}(T(y^j) - h(x^j))x_i^j$ below:

$$\frac{\partial ln(\mathcal{L}(\theta))}{\partial \theta_i} = \frac{\partial ln(\mathcal{L}(\theta))}{\partial \eta}\frac{\partial \eta}{\partial \theta_i}$$

$$\frac{\partial ln(\mathcal{L}(\theta))}{\partial \eta} = \frac{\partial}{\partial \eta}\sum_{j=1}^{m} ln(b(y^j)e^{\eta T(y^j) - A(\eta)})$$

$$= \frac{\partial}{\partial \eta}\sum_{j=1}^{m} ln(b(y^j)) + ln(e^{\eta T(y^j) - A(\eta)})$$

$$= \sum_{j=1}^{m}\frac{\partial}{\partial \eta}\eta T(y^j) - A(\eta)$$

$$= \sum_{j=1}^{m} T(y^j) - \frac{\partial A(\eta)}{\partial \eta}$$

$$= \sum_{j=1}^{m} T(y^j) - h(x)$$

$$\frac{\partial ln(\mathcal{L}(\theta))}{\partial \eta}\frac{\partial \eta}{\partial \theta_i} = \sum_{j=1}^{m}(T(y^j) - h(x))\frac{\partial \eta}{\partial \theta_i}$$

$$\frac{\partial ln(\mathcal{L}(\theta))}{\partial \eta}\frac{\partial \eta}{\partial \theta_i} = \sum_{j=1}^{m}(T(y^j) - h(x))x_i^j$$

We can vectorize gradient ascent on the GLM as follows:

$$\theta = \theta + \nabla_\theta ln(\mathcal{L}(\theta))$$

$$\theta = \theta + X^t(Y - h(X))$$

$$\theta = \theta + X^t(Y - A(X\theta))$$

Where $Y \in \mathbb{R}^m, X \in \mathbb{R}^{mx(n+1)}$ and θ is exponential family dependent.

7 Explain It Like I'm 5

GLMs draw a "line" to better understand the data. The line is drawn based upon assumptions between the desired result and the data itself. For example, image you have a room with red and blue balls. You want to classify red balls from blue balls. An easy way to do so would be to draw a line that best separates the two.