Generalized Linear Model (GLM)

Teeradaj Racharak (เอ็กซ์)
r.teeradaj@gmail.com

THAI PROGRAMMER

SOFTWARE PARK THAILAND

Software Park {Digital Academy}

LEAN UPSKILL

Amazing Results (Recap)

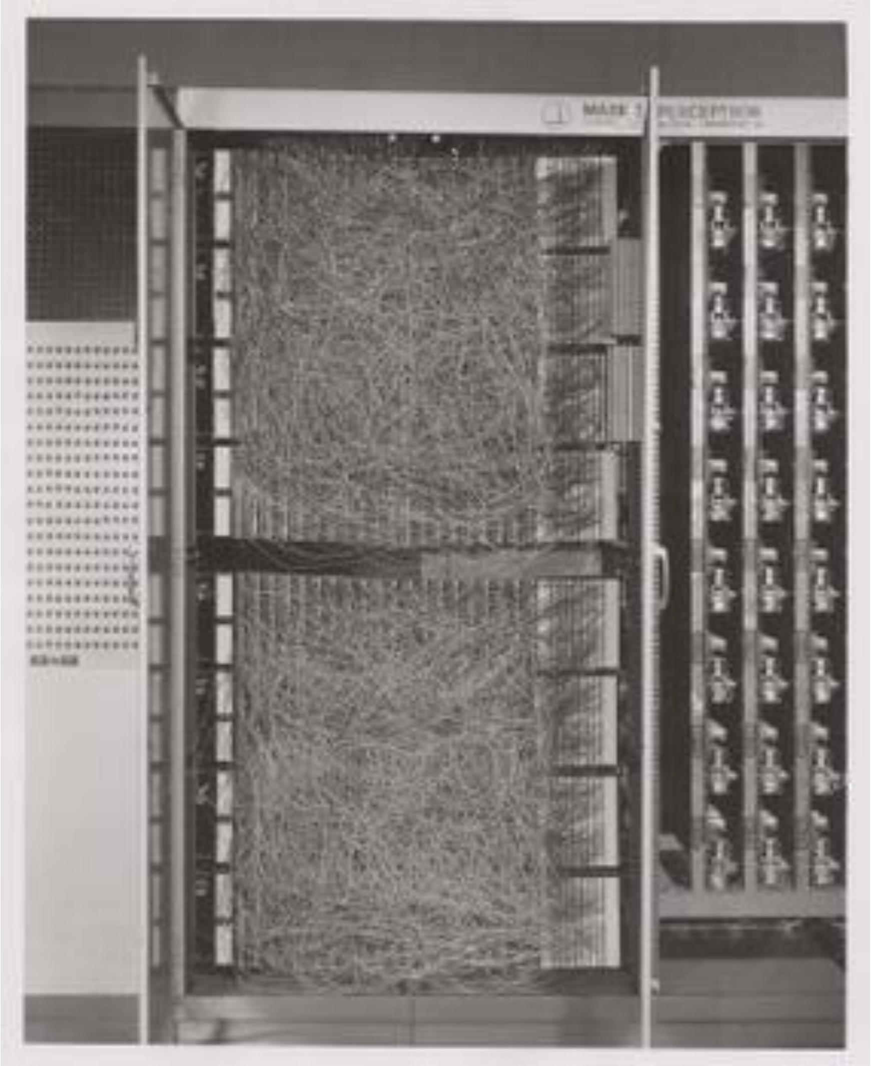
Update rules in iterative methods:

$$\boldsymbol{\theta}^{(n+1)} := \boldsymbol{\theta}^{(n)} + \alpha(y^{(i)} - h_\theta(\boldsymbol{x}^{(i)}))\boldsymbol{x}^{(i)}$$

Although the rules are not exactly the same since $h_\theta(\boldsymbol{x})$ is not the same, it is pretty amazing that we get similar update rules for:

1. Linear regression with least squares
2. Linear regression with maximum likelihood
3. Logistic regression with maximum likelihood
4. Logistic regression with negation of log loss

The 1st Neural Network (Recap)



Mark I Perceptron Machine (Wikipedia)

In 1957, Rosenblatt conceived of the perceptron, a physical machine implementing the classification function

$$h_\theta(\boldsymbol{x}) = g(\boldsymbol{\theta}^T \boldsymbol{x})$$

with

$$g(z) = \begin{cases} 1 & \textbf{if } z \geq 0 \\ 0 & \textbf{otherwise} \end{cases}$$

The 1st Neural Network (Recap)

The perceptron learning algorithm also used the update rule:

$$\boldsymbol{\theta}^{(n+1)} := \boldsymbol{\theta}^{(n)} + \alpha(y^{(i)} - h_\theta(\boldsymbol{x}^{(i)}))\boldsymbol{x}^{(i)}$$

However, this is different from the logistic and linear regression rules since $h_\theta(\boldsymbol{x})$ is in this case a hard threshold classifier without any probabilistic interpretation.

Motivation

Can we find a generalized linear model from what we have observed ?

- Linear regression with least squares
- Linear regression with maximum likelihood
- Logistic regression with maximum likelihood
- Logistic regression with negation of log loss

We'll see later that both linear regression and logistic regression are generalized linear models (GLMs).

- Why is this reasonable?

Motivation

In linear regression, we observe a random variable y assumed to be drawn from a Gaussian distribution depending linearly on a random variable vector $\boldsymbol{x}$ drawn from some population with conditional density.

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^T\boldsymbol{x}, \sigma^2)$$

In linear regression, we observe a random variable y assumed to be drawn from a Gaussian distribution depending linearly on a random variable vector $\boldsymbol{x}$ drawn from some population with conditional density.

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \textbf{Bernoulli}(\boldsymbol{\theta}^T\boldsymbol{x}, \sigma^2)$$

Generalized Linear Models (GLMs)

To understand GLMs, we need to understand the exponential family of distributions. We say that a class of distributions is in the exponential family if it can be rewritten in the form:

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

where

- η is the natural parameter or canonical parameter of the distribution,
- $T(y)$ is the sufficient statistic (we normally use $T(y) = y$),
- $a(\eta)$ is the log partition function (we use $e^{-a(\eta)}$ just to normalize the distribution to have sum or integral of 1), and
- $b(y)$ is an arbitrary scalar function of y.

Each choice of T, a, and b defines a family (set) of distributions parameterized by η .

Generalized Linear Models (GLMs)

The Gaussian and Bernoulli distributions are both exponential family distributions.

If $y = \textsf{\textbf{Bernoulli}}(\phi)$, then $p(y=1;\phi)=\phi$ and $p(y=0;\phi)=1-\phi$.

We thus rewrite the above in a compact form as follows:

$$
\begin{aligned}
p(y;\phi) &= \phi^{y}(1-\phi)^{1-y} \\
&= e^{(y\log\phi+(1-y)\log(1-\phi))} \quad \text{(using the substitution } z = e^{\log z}\text{)} \\
&= e^{(\log\frac{\phi}{1-\phi})y+\log(1-\phi)}
\end{aligned}
$$

That's, $\eta = \log\dfrac{\phi}{1-\phi}$, $T(y)=y$, $a(\eta)=\log(1+e^{\eta})$, and $b(y)=1$.

i.e. we see that $p(y;\phi)$ is in the exponential family.

Generalized Linear Models (GLMs)

The Gaussian and Bernoulli distributions are both exponential family distributions.

If $y = \mathcal{N}(\mu, \sigma^2)$ and assume that $\sigma^2 = 1$, then we have:

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

$$= \frac{1}{2\pi} e^{-\frac{1}{2}y^2} e^{\mu y - \frac{1}{2}\mu^2}$$

That's, $\eta = \mu$, $T(y) = y$, $a(\eta) = \frac{\mu^2}{2} = \frac{\eta^2}{2}$, and $b(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$.

$i.e.$ we see that $p(y;\mu)$ is in the exponential family, too !

Exponential Family

There are other useful members in the exponential family such as:

- Multinomial for *k*-class classification problems
- Poisson for modeling count data
- Gamma and exponential for continuous non-negative random variables like time intervals
- Beta and Dirichlet for distributions over probabilities
- and *etc*.

If y is in the exponential family given $\boldsymbol{x}$ and $\boldsymbol{\theta}$, we can apply the same procedure (*i.e.* the GLM recipe) to come up with a model !

Let's see the recipe in the next slide !

Recipe for Logistic Regression

The GLM makes three assumptions as follows:

1. $p(y|\boldsymbol{x};\boldsymbol{\theta}) =$ **ExponentialFamily**(η) .
2. Given $\boldsymbol{x}$, we would like to predict an expected value of $T(y)$ given $\boldsymbol{x}$.
3. η is linear *i.e.* $\eta_i := \boldsymbol{\theta}_i^T\boldsymbol{x}$.

Assumption 2 means that we want to learn a hypothesis function $h(\boldsymbol{x}) = E[y|\boldsymbol{x}]$

For logistic regression, this would be:

$$\begin{aligned} h_{\boldsymbol{\theta}}(\boldsymbol{x}) &= E[y|\boldsymbol{x};\boldsymbol{\theta}] \\ &= 0 \cdot p(y=0|\boldsymbol{x};\boldsymbol{\theta}) + 1 \cdot p(y=1|\boldsymbol{x};\boldsymbol{\theta}) \\ &= 1 \cdot p(y=1|\boldsymbol{x};\boldsymbol{\theta}) \end{aligned}$$

Recipe for Linear Regression

In the linear regression setting, if we apply the GLM assumptions with the Gaussian distribution, we then obtain:

$$y \sim \mathcal{N}(\mu, \sigma^2)$$

and $h_{\boldsymbol{\theta}}(\boldsymbol{x})$ needs to be a prediction of $T(y)$ given $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

We have already found that the Gaussian is an exponential family distribution with natural parameter $\eta = \mu$.

Since $\eta = \boldsymbol{\theta}^T \boldsymbol{x}$ and let $T(y) = y$, we obtain:

$$\begin{aligned} h_{\boldsymbol{\theta}}(\boldsymbol{x}) &= E[y \,|\, \boldsymbol{x}; \boldsymbol{\theta}] \\ &= \mu \\ &= \eta \ = \boldsymbol{\theta}^T \boldsymbol{x} \end{aligned}$$

Revisiting Logistic Regression

Now, let's consider the logistic regression setting *i.e.*
we have two classes namely 0 and 1.

If we assume $y \sim$ **Bernoulli**(ϕ) and follow the GLM recipe *i.e.* recasting the Bernoulli distribution as a member of the exponential family, we thus obtain:

$$\phi = \frac{1}{1 + e^{-\eta}}$$

We then try to predict the expectation of $T(y) = y$ given $\boldsymbol{x}$ *i.e.*

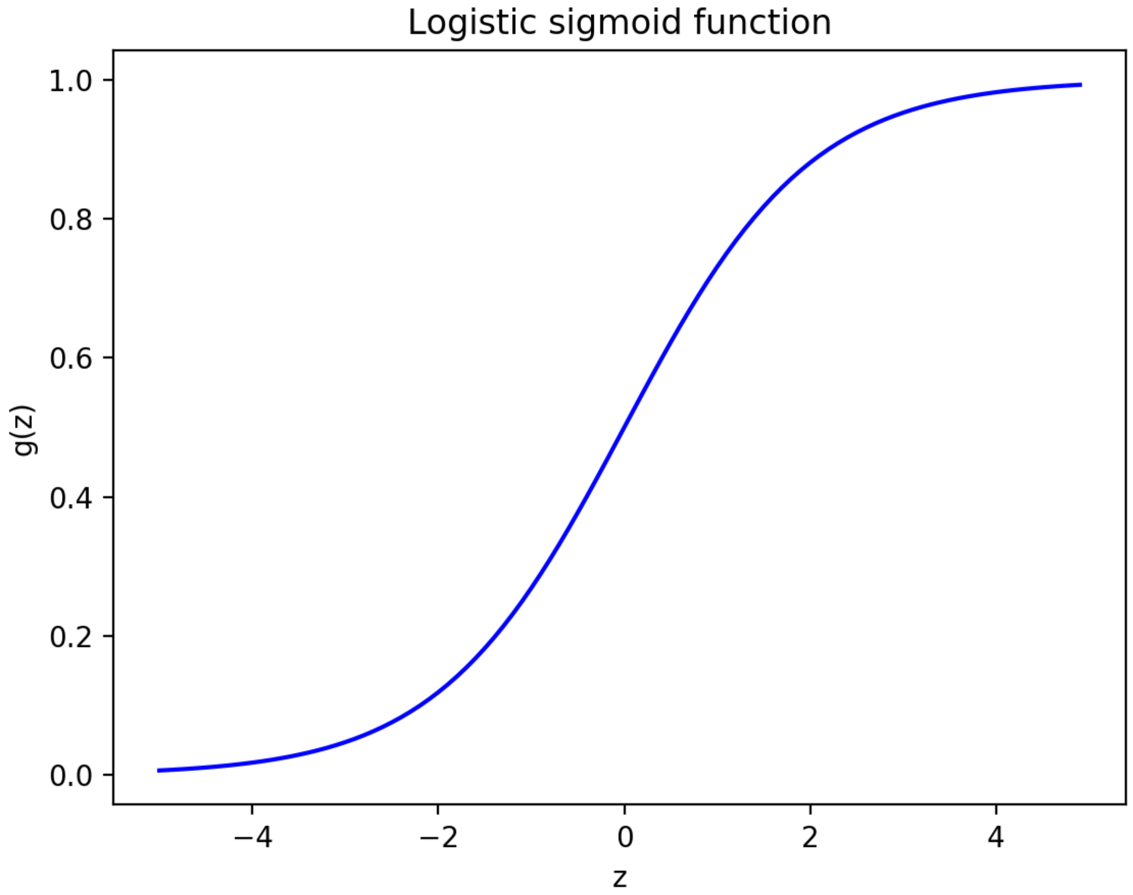
$$\begin{aligned} h_{\boldsymbol{\theta}}(\boldsymbol{x}) &= E[y \mid \boldsymbol{x}; \boldsymbol{\theta}] \\ &= \phi \\ &= \frac{1}{1 + e^{-\eta}} \quad \text{(from the above)} \quad = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}} \end{aligned}$$

Revisiting Logistic Regression

Now, we understand why we should take the logistic sigmoid *i.e.*

$$\frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}$$

as a model for $p(y = 1 \,|\, \boldsymbol{x}; \boldsymbol{\theta})$ in logistic regression.



That's, the logistic sigmoid is the natural consequence of choosing a GLM to model y as a Bernoulli random variable depending on $\boldsymbol{x}$.

The GLM Recipe

We have already known how to do linear regression and logistic regression. So, why should we care about the GLM recipe ?

The reason is that the GLM recipe can be applied to any distribution and usually leads to elegant learning rules.

So, if you have faced with an unseen learning problem, your baseline approach should be:

1. Come up with a model for the conditional distribution of y given $\boldsymbol{x}$.
2. Cast that conditional distribution as a member of the exponential family to determine what η is.
3. Replace η_i with $\boldsymbol{\theta}_i^T \boldsymbol{x}$.
4. Come up with a procedure to maximize $l(\boldsymbol{\theta})$ for a training set.

GLM: Example in Multinomial Distribution

ใช้ classify ปัญหาที่มากกว่า 2 class

As an example, let’s consider the generalization of the logistic regression problem to k classes *i.e.* $\mathcal{Y} = \{1,2,3,\ldots,k\}$

The natural generalization of the Bernoulli distribution is the multinomial distribution with parameters $\phi_1, \ldots, \phi_{k-1}$.

[We leave out the redundant $\phi_k = 1 - \Sigma_{i=1}^{k-1} \phi_i$]

To model the multinomial as a member of the exponential family, there is a rather involved derivation (see [1] for details).

[1] Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

GLM: Example in Multinomial Distribution

The upshot: we can obtain $\eta_i = \log \frac{\phi_i}{\phi_k}$, which can be inverted to obtain:

$$\phi_i = \frac{e^{\eta_i}}{\Sigma_{j=1}^{k} e^{\eta_j}}$$

which is called the softmax function and is the multi-class generalization of the logistic sigmoid.

Our prediction then becomes:

hypothesis fn : softmax

$$p(y = i \mid \boldsymbol{x}; \boldsymbol{\Theta}) = \phi_i = \frac{e^{\eta_i}}{\Sigma_{j=1}^{k} e^{\eta_j}} = \frac{e^{\boldsymbol{\theta}_i^T \boldsymbol{x}}}{\Sigma_{j=1}^{k} e^{\boldsymbol{\theta}_j^T \boldsymbol{x}}}$$

See [1] for the log likelihood function and the applications of gradient methods to find the optimal $\boldsymbol{\Theta}$.

[1] Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

GLM: Example in Multinomial Distribution

Lastly, let's discuss parameter fitting. If we have a training set of m examples and would like to learn the parameters θ_i of this model, we would write down the log likelihood:

$$l(\theta) = \Sigma_{i=1}^{m} \log p(y^{(i)} | x^{(i)}; \theta)$$

$$= \Sigma_{i=1}^{m} \log \Pi_{l=1}^{k} \left(\frac{e^{\boldsymbol{\theta}_l^T \boldsymbol{x}^{(i)}}}{\Sigma_{j=1}^{k} e^{\boldsymbol{\theta}_j^T \boldsymbol{x}^{(i)}}} \right)^{1\{y^{(i)}=l\}}$$

* ถ้า {} เป็นจริงจะได้ 1{...} = 1
ถ้าไม่จริง 1{...} = 0

We can now obtain the maximum likelihood estimate of the parameters by maximizing $l(\theta)$ in terms of θ.

($1\{\cdot\}$ is an indicator function which yields 1 if its argument is true and 0 otherwise)

Summary

To summarize the GLM approach:

- **Assumption 1:** the distribution $p(y|\boldsymbol{x};\boldsymbol{\theta})$ is a member of the exponential family with natural parameter(s) η .
- **Assumption 2:** our goal is to predict the expectation $E[T(y)|\boldsymbol{x};\boldsymbol{\theta}]$ if an input $\boldsymbol{x}$. T is a transformation of y that comes from modeling $p(y|\boldsymbol{x};\boldsymbol{\theta})$ as a member of the exponential family.
- **Assumption 3:** the natural parameter(s) of the distribution η are linear in $\boldsymbol{x}$ *i.e.* $\eta_i := \boldsymbol{\theta}_i^T \boldsymbol{x}$.

If you are willing to make these assumptions, you end up with a 'concave' log likelihood function, which means that any local maximum is a global maximum.

A GLM should be a good first thing to try when you are faced with a machine learning problem you don't already have an algorithm for !

Summary

Let’s revisit our reminder !

1. If you have continuous $\mathscr{X}$ and want to predict continuous $\mathscr{Y}$, then your first go-to model is linear regression !
   - You may also consider non-linear transformation.
2. If you have continuous $\mathscr{X}$ and want to predict discrete $\mathscr{Y}$, then your first go-to model is logistic or softmax regression, or you can come up with a new GLM from scratch !