Foundation of Analytics: Lecture 4

Dihui Lai

dlai@wustl.edu

August 2, 2020

Logistic Regression: Likelihood Function

Assuming two possible outcomes 1 and 0, the probability of being 1 is modeled as

$$p_i = \frac{1}{1 + \exp(-\vec{\beta} \cdot \vec{x}^i)}$$

The likelihood function is defined as

$$Likelihood = \prod_{i=1}^{n} p_i^{y^i} (1 - p_i)^{1-y^i}$$

The log-likelihood function is the defined as the log transformation of the likelihood function

$$\ell = \log(Likelihood) = \sum_{i=1}^{n} y^i \log(p_i) + (1 - y^i) \log(1 - p_i)$$

Logistic Regression: Optimization Attempt

It follows that

$$\begin{aligned}\ell &= \sum_{i=1}^{n} y^i \log \frac{p_i}{1-p_i} + \log(1-p_i) \\ &= \sum_{i=1}^{n} y^i(\vec{\beta}\cdot\vec{x^i}) - \log(1+\exp(\vec{\beta}\cdot\vec{x^i}))\end{aligned}$$

Take the gradient against βs, we have

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n}\left(y^i - \frac{1}{1+\exp(-\vec{\beta}\cdot\vec{x^i})}\right)x_j^i, j = 1,2,3,...,m$$

βs can NOT be solved by setting $\nabla\ell = 0$ because of the nonlinear term of x^i, which is $\frac{1}{1+\exp(\vec{x^i}\cdot\vec{\beta})}$.

Newton-Raphson Method for Optimizing Non-linear Functions

Consider a function of one parameter $\ell(\beta)$ and assume β_0 is close to the point that minimizes $\ell(\beta)$. We can therefore use Talyor expansion for approximation

$$\ell(\beta) = \ell(\beta_0) + \ell'(\beta_0)(\beta - \beta_0) + \frac{1}{2}\ell''(\beta_0)(\beta - \beta_0)^2$$

The β^* that minimize the function have derivative at the point 0 i.e. $\ell'(\beta)|_{\beta=\beta^*} = 0$, by setting $\ell'(\beta) = 0$, we get an iterative evaluation methods for β^*

$$\ell'(\beta_0) + \frac{1}{2}2\ell''(\beta_0)(\beta - \beta_0) = 0 \rightarrow \beta = \beta_0 - \frac{\ell'(\beta_0)}{\ell''(\beta_0)} \text{i.e.}$$

$$\beta^{(k+1)} = \beta^{(k)} - \frac{\ell'(\beta^{(k)})}{\ell''(\beta^{(k)})}$$

Multivariate Newton-Raphson Method

For multivarite function, the iteration formula becomes

$$\beta^{(k+1)} = \beta^{(k)} - H^{-1}(\beta^{(k)})\nabla\ell(\beta^{(k)})$$

here $H(\beta^{(k)})$ is the Hessian matrix of $\ell(\beta)$ evaluated at $\beta = \beta^{(k)}$, defined as

$$H_{ab} = \frac{\partial^2 \ell}{\partial\beta_a \partial\beta_b}|_{\beta=\beta^{(k)}}$$

and $H^{-1}(\beta^{(k)})$ is the inverse of $H(\beta^{(k)})$

Logistic Regression

Apply Newton-Raphson methods to optimize the logistic regression, we calculate the Hessian of the log-likelihood function

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \beta_a \partial \beta_b} &= -\sum_{i=1}^{n} x_b^i \frac{\exp(-\vec{\beta}\cdot\vec{x}^i)}{(1+\exp(-\vec{\beta}\cdot\vec{x}^i))^2} x_a^i \\ &= -\sum_{i=1}^{n} x_b^i p_i(1-p_i) x_a^i\end{aligned}$$

written in matrix formula, the Hessian of the loglikelihood function is

$$H = -X^T W X,\ W = \begin{bmatrix} p_1(1-p_1) & & \\ & \ddots & \\ & & p_n(1-p_n) \end{bmatrix}$$

Logistic Regression: Optimization Algorithm

Use Newton Raphson Methods, we have

$$\vec{\beta}^{(k+1)} \leftarrow \vec{\beta}^{(k)} - H^{-1}\nabla\ell$$
$$\vec{\beta}^{(k+1)} \leftarrow \vec{\beta}^{(k)} + (X^T W X)^{-1} X^T (y - p)$$

Recall in linear regression case

$$\beta = (X^T X)^{-1} X^T y$$

Generalized Linear Model

Exponential family of probability density function

$$f(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

The distribution have the following properties

- $E(Y) = b'(\theta)$
- $Var(Y) = b''(\theta)a(\phi)$

Generalized Linear Model: Gaussian

Gaussian distribution as a special case of exponential family

$$f(y) = \exp\left(\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)$$

where we have $a(\phi) = \sigma^2$, $b(\mu) = \frac{1}{2}\mu^2$ Therefore

- $E(Y) = b'(\theta) = \mu$
- $Var(Y) = b''(\theta)a(\phi) = \sigma^2$

Link Function

Assume a linear model where

$$\eta = \vec{x} \cdot \vec{\beta}$$
$$b'(\theta) = g(\eta) = g(\vec{x} \cdot \vec{\beta})$$
$$b''(\theta)\frac{\partial \theta}{\partial \beta_j} = V(\mu)\frac{\partial \theta}{\partial \beta_j} = g'(x)x_j$$

here $g^{-1}(\cdot)$ is called the link function. $b''(\theta) = V(\mu)$

Log Likelihood Function of GLM

The log likelihood function of GLM

$$\ell = \sum_i \frac{y^i\theta^i - b(\theta^i)}{a(\phi)} + c^i(y^i, \phi)$$

In the model, only θ is depedent on $\vec{x} \cdot \vec{\beta}$. Therefore, "maximize" the likelihood function is equivalent to minimize

$$\ell = -2\sum_i \left[y^i\theta^i - b(\theta^i)\right]$$

Log Likelihood Function of GLM

Using Newton-Raphson method

$$\beta^{(m+1)} = \beta^{(m)} - H^{-1}(\beta^{(m)})\nabla\ell(\beta^{(m)})$$

We need to calculate the gradient of ℓ and its Hessian

The Gradient and Hessian

The gradient can be derived as

$$\frac{\partial \ell}{\partial \beta_j} = -2\sum_i (y^i - b'(\theta^i))\frac{\partial \theta^i}{\partial \beta_j}$$

$$\frac{\partial \ell}{\partial \beta_j} = -2\sum_i (y^i - \mu^i)\frac{g'(\eta^i)}{V(\mu^i)}x_j^i$$

The hessian can be derived as

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j} = 2\sum_i \left[\frac{g'(\eta^i)^2}{V(\mu^i)} - (y^i - \mu^i)\frac{g''(\eta^i)V(\mu^i) - g'(\eta^i)^2 V'(\mu^i)}{V(\mu^i)^2}\right] x_j^i x_k^i$$

Optimization: Gradient Descent Method

Cost function $J(\beta)$
Update methods

$$\beta_j \leftarrow \beta_j - \epsilon \frac{\partial}{\partial \beta_j} J(\beta)$$

where ϵ is the learning rate

Gradient Descent Method for Linear Regression

Cost function $J(\beta) = \sum_i \frac{1}{2}(y^i - \vec{x}^i \cdot \vec{\beta})^2$

$$\frac{\partial J}{\partial \beta_j} = \sum_i (\hat{y}^i - y^i) x_j^i$$

Update methods is now

$$\beta_j \leftarrow \beta_j + \epsilon (y^i - \hat{y}^i) x_j^i$$

The update method is quite intuitive considering that β_i is adjusted higher if estimated $\hat{y}^j$ is less than y^j; adjusted lower if $\hat{y}^j$ is more than y^j

Batch/Stochastic Gradient Descent

Batch Gradient Descent: if each step β_i is updated using all data points

$$\beta_j \leftarrow \beta_j + \sum_i \epsilon \frac{\partial}{\partial \beta_j} J(\beta)$$

or

$$\beta_j \leftarrow \beta_j + \sum_i \epsilon (y^i - \hat{y}^i) x_j^i$$

Stochastic Gradient Descent: if each step β_i is updated using only one data point

$$\beta_j \leftarrow \beta_j + \epsilon \frac{\partial}{\partial \beta_j} J(\beta)$$

or

$$\beta_j \leftarrow \beta_j + \epsilon (y^i - \hat{y}^i) x_j^i$$