Foundation of Analytics: Lecture 4

Dihui Lai

dlai@wustl.edu

March 29, 2020

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

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