Foundation of Analytics: Lecture 4

Dihui Lai

dlai@wustl.edu

August 18, 2019

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

$$\theta = \eta = \vec{x} \cdot \vec{\beta}$$
$$b'(\theta) = \mu = g(\vec{x} \cdot \vec{\beta})$$

here $g^{-1}(\cdot)$ is the link function

Log Likelihood Function of GLM

The likelihood function of GLM

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

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

$$\ell = \sum_j \left[y^j \theta^j - b(\theta^j) \right]$$

Log Likelihood Function of GLM

Let's consider each data point and its contribution to the likelihood function

$$\ell^j = y^j\theta^j - b(\theta^j)$$

or simplified as

$$\ell = y\theta - b(\theta)$$

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

$$\frac{\partial \ell}{\partial \beta_i} = -2 \sum_j (y^j - \mu^j) \frac{g'(\eta^j)}{V(\mu)^j} x_i^j$$

The hessian can be derived as

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_i} = 2 \sum_j \left[\frac{g'(\eta^j)^2}{V(\mu^j)} - (y^j - \mu^j) \frac{g''(\eta^j) V(\mu^j) - g'(\eta^j)^2 V'(\mu^j)}{V(\mu^j)^2} \right] x_i^j x_k^j$$

Optimization: Gradient Descent Method

Cost function $J(\beta)$
Update methods

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

where ϵ is the learning rate

Gradient Descent Method for Linear Regression

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

$$\frac{\partial J}{\partial \beta_i} = \sum_j (\hat{y}^j - y^j) x_i^j$$

Update methods is now

$$\beta_i \leftarrow \beta_i + \epsilon (y^j - \hat{y}^j) x_i^j$$

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

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

or

$$\beta_i \leftarrow \beta_i + \sum_j \epsilon (y^j - \hat{y}^j) x_i^j$$

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

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

or

$$\beta_i \leftarrow \beta_i + \epsilon (y^j - \hat{y}^j) x_i^j$$