Linear Model (LM) and Generalized Linear Model (GLM)

Dihui Lai

dlai@wustl.edu

May 4, 2019

Content

- Random Variables: Dependent, Independent, Correlation
- Linear Regression of One Variable
- Linear Regression of Multiple Variables
- Logistic Regression

Relationships between Random Variables

Let's look at a few pairs of data points?

- $\mathbf{X} = [0.5, 0.6, 0.1, -0.3, 2.3], \mathbf{Y} = [0.5, 0.6, 0.1, -0.3, 2.3]$
- $\mathbf{X} = [0.5, 0.6, 0.1, -0.3, 2.3], \mathbf{Y} = [0.6, 0.6, 0.12, -0.3, 2.3]$
- $\mathbf{X} = [0.5, 0.6, 0.1, -0.3, 2.3], \mathbf{Y} = [0.02, -0.2, 0.2, 2.1, -0.5]$

What can you tell about the relationship between **X** and **Y**?

Relationships between Random Variables

Given two random variables X and Y, denote the mean and variance of the two variables as $E[X] = \mu_X$, $E[Y] = \mu_Y$, $Var[X] = \sigma_X^2$, $Var[Y] = \sigma_Y^2$.
The covariance of X and Y is the number defined by

$$\begin{aligned} Cov(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY] - \mu_X \mu_Y \end{aligned}$$

The correlation of the two random variables is the number defined by

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Relationships between Random Variables

Calculate the covariance/correlation of

- Example 1:

$$\mathbf{X} = [2, -2, -2, 2], \mathbf{Y} = [2, -2, -2, 2]$$

  We have $\mu_X = 0$, $\mu_Y = 0$, $\sigma_X^2 = 4$, $\sigma_Y^2 = 4$, $E[XY] = 4$ Therefore $Cov(X, Y) = 4 - 0 = 4$ and $\rho_{XY} = 4/(2 * 2) = 1$

- Example 2:

$$\mathbf{X} = [2, -2, -2, 2], \mathbf{Y} = [2, 0, -2, 0]$$

  We have $\mu_X = 0$, $\mu_Y = 0$, $\sigma_X^2 = 4$, $\sigma_Y^2 = 2$, $E[XY] = 2$ Therefore $Cov(X, Y) = 2 - 0 = 2$ and $\rho_{XY} = 2/(2 * \sqrt{2}) = 1/\sqrt{2}$

Linear Regression with One Variable

Data set:

$$\mathbf{Y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{bmatrix}$$

Linear Regression with One Variable

Assume y is linearly depending on x i.e.

$$y = \beta_0 + \beta_1 x$$

Find $\hat{\beta}$ that minimize the estimation error

$$J = \sum_{j=1}^{n} (y^{(j)} - \hat{y}^{(j)})^2 = \sum_{j=1}^{n} (y^{(j)} - \hat{\beta}_0 - \hat{\beta}_1 x^{(j)})^2$$

i.e.

$$\frac{\partial J}{\partial \hat{\beta}_1} = 0 \rightarrow \sum_{j=1}^{n} (y^{(j)} - \hat{\beta}_0 - \hat{\beta}_1 x^{(j)}) x^{(j)} = 0$$

$$\frac{\partial J}{\partial \hat{\beta}_0} = 0 \rightarrow \sum_{j=1}^{n} (y^{(j)} - \hat{\beta}_0 - \hat{\beta}_1 x^{(j)}) = 0$$

$$\hat{\beta}_0 \sum_{j=1}^{n} x^{(j)} = \sum_{j=1}^{n} y^{(j)} x^{(j)} - \hat{\beta}_1 \sum_{j=1}^{n} x^{(j)} x^{(j)}$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{j=1}^{n} (y^{(j)} - \hat{\beta}_1 x^{(j)}) = \bar{y} - \hat{\beta}_1 \bar{x}$$

Insert the second equation to the first, we have

$$n\bar{x}\bar{y} - \bar{\beta}_1 n\bar{x}\bar{x} = \sum_{j=1}^{n} y^{(j)} x^{(j)} - \hat{\beta}_1 \sum_{j=1}^{n} x^{(j)} x^{(j)}$$

Therefore,

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_{j=1}^{n} y^{(j)} x^{(j)} - \bar{x}\bar{y}}{\frac{1}{n} \sum_{j=1}^{n} x^{(j)} x^{(j)} - \bar{x}^2} = \frac{Cov(X, Y)}{Var(X)} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$$

Multivariate Linear Regression

Data set:

$$
\begin{bmatrix}\mathbf{Y},\mathbf{X}\end{bmatrix} = \begin{bmatrix} y^{(1)} & x_0^{(1)} & x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ y^{(2)} & x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y^{(n)} & x_0^{(n)} & x_1^{(n)} & x_2^{(n)} & \dots & x_p^{(n)} \end{bmatrix}
$$

or explicitly

$$
\mathbf{Y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^{(n)} & x_1^{(n)} & x_2^{(n)} & \dots & x_p^{(n)} \end{bmatrix}
$$

Multivariate Linear Regression

Assume y is a linear superposition of multiple x's

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

or simply

$$y = \sum_{i=1}^{p} \beta_i x_i$$

Estimate $\hat{\beta}$'s that best fits the data i.e. For each estimated data point

$$\hat{y}^{(j)} = \sum_{i=1}^{p} \hat{\beta}_i x_i^{(j)}$$

we need to minimize the error

$$J(\beta) = \sum_{j=1}^{n} (y^{(j)} - \hat{y}^{(j)})^2$$

Multivariate Linear Regression

Solve the optimization problem in matrix format

$$\frac{\partial J}{\partial \beta_i} = 0$$

i.e.

$$\sum_{j=1}^{n} \frac{\partial (y^{(j)} - \hat{y}^{(j)})^2}{\partial \beta_i} = 0$$

$$\sum_{j=1}^{n} (y^{(j)} - \hat{y}^{(j)}) \frac{\partial \hat{y}^{(j)}}{\partial \beta_i} = 0$$

$$\sum_{j=1}^{n} (y^{(j)} - \hat{y}^{(j)}) x_i^{(j)} = 0$$

Multivariate Linear Regression

written in matrix formula we have

$$J(\beta) = (\mathbf{Y} - \mathbf{X}\hat{\beta})^T \mathbf{X}$$

$$\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\hat{\beta} = 0$$

Therefore

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Logistic Regression

$$
\begin{aligned}
p &= \frac{1}{1+\exp(-\vec{\beta}\cdot\vec{x}^j)} \\
L &= \prod_{j=1}^{n} p_j^{y_j}(1-p_j)^{1-y_j} \\
\ell &= \log(L) \\
&= \sum_{j=1}^{n} y_j \log(p_j) + (1-y_j)\log(1-p_j) \\
&= \sum_{j=1}^{n} y_j \log\frac{p_j}{1-p_j} + \log(1-p_j) \\
&= \sum_{j=1}^{n} y^j(\vec{\beta}\cdot\vec{x}^j) - \log(1+\exp(\vec{\beta}\cdot\vec{x}^j)))
\end{aligned}
$$

Logistic Regression

$$
\begin{aligned}
\frac{\partial \ell}{\partial \beta_i} &= \sum_{j=1}^{n} y^j x_i^j - \frac{x_i^j}{1+\exp(-\vec{\beta}\cdot\vec{x}^j)} \\
&= \sum_{j=1}^{n} \left(y^j - \frac{1}{1+\exp(-\vec{\beta}\cdot\vec{x}^j)} \right) x_i^j \\
&= \sum_{j=1}^{n} \left(y^j - \hat{y}^j \right) x_i^j \\
\nabla \ell &= \mathbf{X}^T (\mathbf{Y} - \hat{\mathbf{Y}})
\end{aligned}
$$

βs can not be solved by setting $\nabla \ell = 0$ because of the nonlinear formula for $\hat{y}^j = \frac{1}{1+\exp(\vec{x}^j \cdot \vec{\beta}^j)}$. Recall $\hat{y}^j = \vec{x}^j \cdot \vec{\beta}^j$ for linear regression.

Newton-Raphson Method for Optimizatoin

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

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

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

$$f'(\beta_0) + \frac{1}{2}2f''(\beta_0)(\beta - \beta_0) = 0$$

$$\rightarrow \beta = \beta_0 - \frac{f'(\beta_0)}{f''(\beta_0)} \text{i.e.}$$

$$\beta^{(m+1)} = \beta^{(m)} - \frac{f'(\beta^{(m)})}{f''(\beta^{(m)})}$$

Multivariate Newton-Raphson Method

For multivarite function, the iteration formula becomes

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

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

$$H_{ij} = \frac{\partial^2 f}{\partial \beta_i \partial \beta_j}|_{\beta=\beta^{(m)}}$$

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

Logistic Regression

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

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_i} &= -\sum_{j=1}^{n} x_i^j \frac{\exp(-\vec{\beta} \cdot \vec{x}^j)}{(1+\exp(-\vec{\beta} \cdot \vec{x}^j))^2} x_k^j \\ &= -\sum_{j=1}^{n} x_i^j \hat{y}^j (1-\hat{y}^j) x_k^j\end{aligned}$$

written in matrix formula

$$\mathbf{H} = -\mathbf{X}^T \mathbf{W} \mathbf{X},\ \mathbf{W} = \begin{bmatrix} \hat{y}^1(1-\hat{y}^1) & & \\ & \ddots & \\ & & \hat{y}^n(1-\hat{y}^n) \end{bmatrix}$$

Logistic Regression

$$\vec{\beta}^{(m+1)} \leftarrow \vec{\beta}^{(m)} - \mathbf{H}^{-1}\nabla\ell$$

$$\vec{\beta}^{(m+1)} \leftarrow \vec{\beta}^{(m)} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{Y}})$$

by defining $z^{(m)} = \mathbf{X}\beta^{(m)} + \mathbf{W}^{-1}(\mathbf{Y} - \hat{\mathbf{Y}})$ we have

$$\vec{\beta}^{(m+1)} \leftarrow (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}z^{(m)}$$

Recall in linear regression case

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

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

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

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

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