Foundation of Analytics: Lecture 3

Dihui Lai

dlai@wustl.edu

March 9, 2020

Content

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

Multivariate Linear Regression

Data set:

$$\left[y, X\right] = \begin{bmatrix} y^1 & x_0^1 & x_1^1 & x_2^1 & \dots & x_m^1 \\ y^2 & x_0^2 & x_1^2 & x_2^2 & \dots & x_m^2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y^n & x_0^n & x_1^n & x_2^n & \dots & x_m^n \end{bmatrix}$$

Multivariate Linear Regression

Assume y is a linear superposition of multiple x's

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

or simply

$$\hat{y} = \sum_{j=1}^{m} \beta_j x_j$$

Multivariate Linear Regression

Estimate β's that best fits the data, we need to minimize the error

$$\epsilon = \sum_{i=1}^{n}(y^i - \hat{y}^i)^2$$
$$= (y - \hat{y})^T(y - \hat{y})$$

Use basic calculus we know, we want to have the βs satisfy the following equation set:

$$\frac{\partial \epsilon}{\partial \beta_j} = 0, j = 1, 2, 3, 4...m$$

i.e.

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

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

$$\sum_{i=1}^{n} (y^i - \hat{y}^i) x_j^i = 0$$

Multivariate Linear Regression

Written in matrix formula we require

$$(y - X\beta)^T X = 0$$

or after transposing

$$X^T y - X^T X \beta = 0$$

Therefore

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

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

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

$$= \sum_{i=1}^{n} y^i(\vec{\beta} \cdot \vec{x}^i) - \log(1 + \exp(\vec{\beta} \cdot \vec{x}^i)))$$

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

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

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

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

$$= -\sum_{i=1}^{n} x_b^i p_i (1 - p_i) x_a^i$$

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

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

Recall in linear regression case

$$\beta = (X^TX)^{-1}X^Ty$$