Foundation of Analytics: Lecture 3

Dihui Lai

dlai@wustl.edu

March 29, 2020

Content

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

Correlations between Random Variables

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

- $\vec{x} = [0.5, 0.6, 0.1, -0.3, 2.3], \vec{y} = [0.5, 0.6, 0.1, -0.3, 2.3]$
- $\vec{x} = [0.5, 0.6, 0.1, -0.3, 2.3], \vec{y} = [0.6, 0.6, 0.12, -0.3, 2.3]$
- $\vec{x} = [0.5, 0.6, 0.1, -0.3, 2.3], \vec{y} = [0.02, -0.2, 0.2, 2.1, -0.5]$

What can you tell about the relationship between $\vec{x}$ and $\vec{x}$?

Correlations 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

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

Empricial Estimation of Covariance

$$Cov(X,Y) = \frac{(x-\mu_x)(y^T-\mu_y)}{N} (\textit{empirical})$$

$$Cov(X,Y) = \frac{(x-\mu_x)(y^T-\mu_y)}{N-1} (\textit{unbiased})$$

Correlations between Random Variables

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

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

Correlations between Random Variables

Calculate the covariance/correlation of

- Example 1:
$$\vec{x} = [2, -2, -2, 2], \vec{y} = [2, -2, -2, 2]$$
We have $\mu_x = 0$, $\mu_y = 0$, σ_x^2 =4, σ_y^2 =4, $E[XY]$ =4 Therefore $Cov(X, Y) = 4 - 0 = 4$ and $\rho_{xy} = 4/(2 * 2) = 1$
- Example 2:
$$\vec{x} = [2, -2, -2, 2], \vec{y} = [2, 0, -2, 0]$$
We have $\mu_x = 0$, $\mu_y = 0$, σ_x^2 =4, σ_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:

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

Linear Regression with One Variable

Assume y is linearly depending on x i.e.

$$\hat{y} = \beta_0 + \beta_1 x$$

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

$$\epsilon = \sum_{i=1}^{n} (y^i - \hat{y}^i)^2 = \sum_{i=1}^{n} (y^i - \beta_0 - \beta_1 x^i)^2$$

i.e.

$$\frac{\partial \epsilon}{\partial \beta_1} = 0 \rightarrow \sum_{i=1}^{n} (y^i - \beta_0 - \beta_1 x^i) x^i = 0$$

$$\frac{\partial \epsilon}{\partial \beta_0} = 0 \rightarrow \sum_{i=1}^{n} (y^i - \beta_0 - \beta_1 x^i) = 0$$

$$\beta_0 \sum_{i=1}^{n} x^i = \sum_{i=1}^{n} y^i x^i - \beta_1 \sum_{i=1}^{n} x^i x^i$$

$$\beta_0 = \frac{1}{n} \sum_{i=1}^{n} (y^i - \beta_1 x^i) = \bar{y} - \beta_1 \bar{x}$$

Insert the second equation to the first, we have

$$n\bar{x}\bar{y} - \beta_1 n\bar{x}\bar{x} = \sum_{i=1}^{n} y^i x^i - \beta_1 \sum_{i=1}^{n} x^i x^i$$

Therefore,

$$\beta_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} x^i y^i - \bar{x}\bar{y}}{\frac{1}{n} \sum_{i=1}^{n} x^i x^i - \bar{x}^2} = \frac{Cov(X,Y)}{Var(X)} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$$

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

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

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

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

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

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

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

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

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

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

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

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