

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Multiple Linear Regression

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
 - Use case: When no single attribute alone is good enough to predict the response variable, *i.e., the Pearson correlation between the response and each other variable is low*. However, a linear combination of attributes gives a good prediction.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,
$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$
- Question: *How do we formulate the notion of the best predictor?*

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: *How do we formulate the notion of the best predictor?*

- Answer: The one that minimizes the **Mean Squared Error** (MSE)

Find $(\theta_0, \theta_1, \dots, \theta_m)$ that minimizes $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_i (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

Find $(\theta_0, \theta_1, \dots, \theta_m)$ that minimizes

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:

- $\theta = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$

- $y = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

- $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \theta = X \cdot \theta$

- So, the problem in terms of the above matrices is:

Find θ that minimizes $\frac{1}{n} \|y - X \cdot \theta\|^2$

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

$$\text{Find } (\theta_0, \theta_1, \dots, \theta_m) \text{ that minimizes } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:
 - $\theta = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$
 - $y = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$
 - $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \theta = X \cdot \theta$
 - So, the problem in terms of the above matrices is:
Find θ that minimizes $\frac{1}{n} \|y - X \cdot \theta\|^2$
- How do we solve such an optimization problem?

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

$$\text{Find } (\theta_0, \theta_1, \dots, \theta_m) \text{ that minimizes } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:

$$\theta = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$$

$$y = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \theta = X \cdot \theta$$

- So, the problem in terms of the above matrices is:

$$\text{Find } \theta \text{ that minimizes } \frac{1}{n} \|y - X \cdot \theta\|^2$$

- How do we solve such an optimization problem?
 1. Use calculus to find a close-form solution for θ .
 2. Use numerical hill-climbing technique such as **gradient ascent/descent**.

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

$$\text{Find } (\theta_0, \theta_1, \dots, \theta_m) \text{ that minimizes } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$$

$$\mathbf{y} = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \boldsymbol{\theta} = \mathbf{X} \cdot \boldsymbol{\theta}$$

- So, the problem in terms of the above matrices is:

$$\text{Find } \boldsymbol{\theta} \text{ that minimizes } f(\boldsymbol{\theta}) = \frac{1}{n} \|\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\theta}\|^2$$

- How do we solve such an optimization problem?

1. Use calculus to find a close-form solution for $\boldsymbol{\theta}$.
2. Use numerical hill-climbing technique such as **gradient ascent/descent**.

- Close-form solution:

- $f(\boldsymbol{\theta})$ is a nice convex function. So, it has a unique minima.
- Idea from calculus: Stack up $\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_0}, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_m}$ as a vector and equate it to the $\mathbf{0}$ vector.
- This gives: $(\mathbf{X}^T \mathbf{X}) \cdot \boldsymbol{\theta} = \mathbf{X}^T \mathbf{y}$ which gives
$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Multiple Linear Regression

- Now that we are familiar with using multiple attributes to predict a categorical variable, let us apply the same idea to predict a quantitative variable (called **regression**).
- Using multiple attributes to predict a quantitative variable is called **multiple linear regression**.
- When using attributes $x_1, x_2, \dots, x_m$ to predict y , the prediction is through a linear combination of the predictor variables,

$$\hat{y} \approx \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_m \cdot x_m$$

- Question: How do we formulate the notion of the *best predictor*?
 - Answer: The one that minimizes the **Mean Squared Error** (MSE)

$$\text{Find } (\theta_0, \theta_1, \dots, \theta_m) \text{ that minimizes } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 \cdot x_{1,i} + \dots + \theta_m \cdot x_{m,i}))^2$$

- We can use matrix notation to write the mathematical problem compactly:

$$\theta = (\theta_0, \theta_1, \dots, \theta_m)^T = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_m \end{pmatrix}$$

$$y = (y_1, \dots, y_n)^T = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{m,1} \\ 1 & x_{1,2} & \dots & x_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \dots & x_{m,n} \end{pmatrix} \cdot \theta = X \cdot \theta$$

- So, the problem in terms of the above matrices is:

$$\text{Find } \theta \text{ that minimizes } f(\theta) = \frac{1}{n} \|y - X \cdot \theta\|^2$$

- How do we solve such an optimization problem?
 1. Use calculus to find a close-form solution for θ .
 2. Use numerical hill-climbing technique such as **gradient ascent/descent**.

• Gradient Descent:

- Start with some initial value of θ
- Repeat until convergence:
 - Find the direction along which the function $f(\theta)$ decreases the most. This happens to be $-\Delta f(\theta)$. So, we update $\theta = \theta - \alpha \cdot \Delta f(\theta)$, where α is the **learning rate**. So, $\theta = \theta + (2\alpha/n) \cdot X^T(y - X \cdot \theta)$

End