Linear Regression with Normal Equation

What is Linear Regression and how it works ?

Data science and Machine learning Online Course

Machine Learning - Bahram Jannesar, Soroush Ghaderi

Introduction :

It is not always necessary to run an optimization algorithm to perform linear regression. You can solve a specific algebraic equation — the normal equation — to get the results directly. Although for big datasets it is not even close to being computationally optimal, it's still one of the good options to be aware of.

As a reminder, below there is a linear regression equation in the

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

expanded form.

In a vectorized form it looks like that:

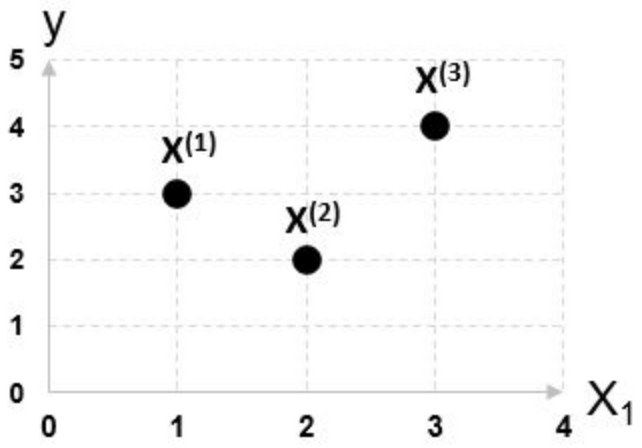
$$\hat{y} = \theta^T \cdot x$$

where θ is a vector of parameters weights.

Usually finding the best model parameters is performed by running some kind of optimization algorithm (e.g. gradient descent) to minimize a cost function. However, it is possible to obtain values (weights) of these parameters by solving an algebraic equation called the normal equation as well. It is defined as below

$$\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot y$$

In this example, there are only three points (observations) with only one variable (X_1). On the graph, they look like below.



In this case, the linear regression equation has a form of:

$$\hat{y} = \theta_0 + \theta_1 x_1$$

Features (X) and labels (y) are:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \qquad y = \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix}$$

Default bias, X_0 — X_1

Note that we add a default bias term of 1 — it will be updated during our calculations. Not adding this term will lead to a wrong solution.

Step 1: Transposition of matrix X

This is a relatively simple task — rows become new columns.

$$X^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$$

Step 2: Multiplication on the transposed matrix and matrix X

$$X^T \cdot X = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$$

Step 3: Inversion of a resultant matrix

To inverse a simple 2x2 matrix we can use the formula:

$$A^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Therefore, we get:

$$(X^T \cdot X)^{-1} = \begin{bmatrix} 2/3 & -1 \\ -1 & 1/2 \end{bmatrix}$$

Step 4: Multiplication of the inverted matrix with X transposed

$$(X^T \cdot X)^{-1} \cdot X^T = \begin{bmatrix} 4/3 & 1/3 & -2/3 \\ -1/3 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$$

$$\begin{bmatrix} 2/3 & -1 \\ -1 & 1/2 \end{bmatrix}$$

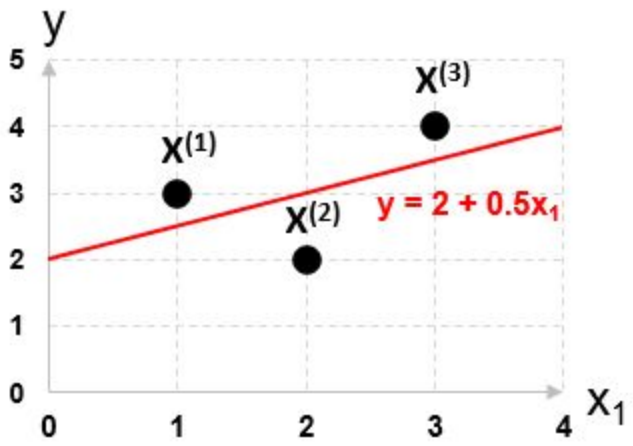
Step 5: Final multiplication to obtain the vector of best parameters

$$\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot y = \begin{bmatrix} 2 \\ 1/2 \end{bmatrix} \begin{bmatrix} 4/3 & 1/3 & -2/3 \\ -1/3 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix}$$

Finally our linear regression equations takes form of:

$$\hat{y} = 2 + 0.5x_1$$

Plotting this line onto a previous graph looks like below.



Gradient Descent Vs Normal Equation:

Gradient Descent

- It requires choosing the value of Alpha.
- It requires many iterations.
- It works well when n (no. of data-set) is large.

Normal Equation

- It does not need to choose the value of Alpha.
- It doesn't require iteration.
- It requires to calculate the inverse of the transpose of x.
- It is slow if n (data-set) if very large.