Introduction

We have known optimization method like gradient descent can be used to minimize the cost function of linear regression. But for linear regression, there exists an analytical solution. That means we can obtain the variables for linear regression in one step calculation by using the right formula. In this post, we will look into the analytical solution of linear regression and its derivations.

Analytical Solution

We first give out the formula of the analytical solution for linear regression. If you are not interested in the derivations, you can just use this formula to calculate your linear regression variables. The solution is:

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

All symbols are vectorized in this formula. If you are not familiar with linear algebra or vectorization, please refer to this blog. In this formula, X is a m by n matrix, which means we have m samples and n feature. The symbol y is a m by 1 vector representing the target label and θ is a n by 1 vector representing all the coefficients we need for each feature.

Derivations

We know the vectorization expression(more details please refer to blog) of linear regression cost function can be denoted as :

$$J(\theta) = \frac{1}{2m}(X\theta - y)^T(X\theta - y)$$

Since 1/(2*m) is a constant, when we minimize a function, multiply or divide the cost function by a non-zero constant doesn't affect the minimization result, thus in this case, we omit this constant term. For convenience, our cost function becomes:

$$J(\theta) = (X\theta - y)^T(X\theta - y)$$

This can be further simplified as:

$$J(\theta) = ((X\theta)^T - y^T)(X\theta - y)$$

We expand it to obtain:

$$J(\theta) = (X\theta)^T(X\theta) - (X\theta)^T y - y^T(X\theta) + y^T y$$

Now need some transformation on the second term. We know X is a m by n matrix and θ is n by 1 matrix, thus Xθ has dimension m by 1 and its transpose has dimension 1 by m. Since y is m by 1, thus the dimension of the second term turns out to be 1. In other words, the second term is a scalar. We know the transpose of a scalar equals to itself, thus we take the transpose of the second term to get:

$$((X\theta)^T y)^T = y^T(X\theta)$$

We substitute is back into our cost function to obtain:

$$J(\theta) = (X\theta)^T(X\theta) - 2y^T(X\theta) + y^T y$$

Further more, we can write it as:

$$J(\theta) = \theta^T X^T X\theta - 2y^T X\theta + y^T y$$

Now we need to take derivative of the cost function. For convenience, the common matrix derivative formulas are listed as reference:

$$\frac{\partial(AX)}{\partial X} = A^T$$

$$\frac{\partial(X^T A)}{\partial X} = A$$

$$\frac{\partial(X^T X)}{\partial X} = 2X$$

$$\frac{\partial(X^T AX)}{\partial X} = AX + A^T X$$

Using the above formulas, we can derive our cost function respect to θ as:

$$\frac{\partial J(\theta)}{\partial \theta} = X^T X\theta + X^T X\theta - 2X^T y = 2X^T X\theta - 2X^T y$$

In order to solve the variables, we need to make the above derivation equal to zero, that is:

$$2X^TX\theta - 2X^Ty = 0$$

We can simplify it as:

$$X^TX\theta = X^Ty$$

Thus we can compute θ as:

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

Conclusion

In this blog, we give the analytical solution of solving the variables for linear regression. We went through the steps of how to derive this result from the derivation of cost function in details.