Deriving the normal equation

Uzair Ahmad

The derivation of the normal equation for multiple linear regression.

Step 1: Define the Problem

In multiple linear regression, you have a dataset with multiple independent variables (features) and one dependent variable (target). You want to find the best-fit linear model that represents the relationship between the features and the target.

Step 2: Formulate the Hypothesis

Assume that the relationship between the features (X) and the target variable (Y) can be modeled as a linear equation:

$$Y = Xw$$

Where:

- Y is the vector of predicted values ($N \times 1$).
- X is the design matrix ($N \times (p+1)$), where N is the number of data points, and p is the number of features. The first column of X is a column of ones for the intercept.
- w is the vector of coefficients (parameters) we want to find (including the intercept), which has dimensions $((p+1) \times 1)$.

Step 3: Define the Cost Function

In linear regression, the cost function is often defined as the mean squared error (MSE), which quantifies the error between the predicted values and the actual target values. The MSE is calculated as follows:

$$J(w) = \frac{1}{2N} \sum_{i=1}^{N} (Y_i - X_i w)^2$$

Where:

- N is the number of data points.
- Y_i is the actual target value for the i^{th} data point.
- X_i is the predicted target value for the i^{th} data point (i.e., the value obtained by applying the linear model).
- w is the vector of coefficients (parameters) we want to optimize.

Now, let's explain why we multiply the error function by 1/2:

1. **Mathematical Convenience:** The factor of 1/2 simplifies the derivative of the cost function with respect to the parameters (w). When you take the derivative of the cost function with this factor, it cancels out the 2 in the exponent, making the subsequent mathematical steps simpler. This doesn't affect the final solution, but it simplifies the mathematics.
2. **Aesthetic and Tradition:** The choice of 1/2 is somewhat arbitrary but has become a common convention in linear regression. It doesn't change the nature of the optimization problem, but it makes the expression for the cost function look neater and more symmetric.

3. **Scaling:** By dividing by 2N, we scale the cost function such that its value is more interpretable and less sensitive to the size of the dataset. This scaling ensures that the cost function doesn't grow excessively with larger datasets, making it easier to compare cost values across different datasets.

In summary, multiplying the error function by 1/2 in the cost function for linear regression is primarily a matter of mathematical convenience, tradition, and scaling to make the cost function more interpretable and robust to dataset size. It doesn't fundamentally change the optimization problem or the resulting coefficients.

Step 4: Minimize the Cost Function

To minimize the cost function, we take the derivative of $J(w)$ with respect to w and set it equal to zero:

$$\nabla J(w) = 0$$

Step 5: Calculate the Gradient $\nabla J(w)$

Calculate the gradient of $J(w)$ with respect to w:

$$\nabla J(w) = -X^T(Y - Xw)$$

Step 6: Set the Gradient Equal to Zero

Set $\nabla J(w)$ equal to zero:

$$-X^T(Y - Xw) = 0$$

Step 7: Solve for w

Solve for w using the normal equation:

$$X^TXw = X^TY$$

To find w, you can use the following formula:

$$w = (X^TX)^{-1}X^TY$$

This equation allows you to calculate the coefficients w that minimize the sum of squared errors and provide the best linear fit to your data with multiple independent variables.

Step 8: Use the Normal Equation Coefficients

Now that you've solved for w, you have the coefficients for your multiple linear regression model. You can use these values to make predictions:

$$Y = Xw$$

This equation represents the best-fit linear model for your data in multiple linear regression.