Data-Driven Problem Solving in Mechanical Engineering

Linear Regression

Masoud Masoumi

ME 371 - Fall 2023

Department of Mechanical Engineering
The Cooper Union for the Advancement of Science and Art

October, 2023

Linear Regression

Given a collection of m points, linear regression seeks to find the line which best approximates or *fits* the points.

There are two main reasons why we want to do this:

- Use it as simplification and compression. We can see the trend and highlight the location and magnitude of outliers

- Use it for value predicting and forecasting.

Linear Regression

Linear Regression seeks the line $y = f(\mathrm{x})$ which minimizes the sum of the squared errors over all points, i.e. the coefficient vector w that minimizes the following squared error function:

$$J(\mathrm{w}) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)} \right)^2$$

with $f(\mathrm{x}) = \omega_o + \sum_{i=1}^{n-1} \omega_i x_i$

where, $\hat{y}^{(i)}$ is the estimated value at $\mathrm{x}^{(i)}$ and m is the number of samples.

Linear Regression can be divided into two types:

- Simple Linear Regression: $f(\mathrm{x}) = \omega_o + \omega_1 x_1$
- Multiple Linear Regression: $f(\mathrm{x}) = \omega_o + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \ldots + \omega_{n-1} x_{n-1}$

Multiple Linear Regression

Multiple linear regression refers to multiple independent variables to make a prediction.

Generally, we seek to find the best values for ω's in
$f(\mathrm{x}) = \omega_o + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \ldots$

$\rightarrow f(\mathrm{x}) = \mathrm{w}^T \mathrm{x}$

$\mathrm{w}^T = [\omega_o, \omega_1, \omega_2, \omega_3, \ldots]$ and $\mathrm{x} = [1, x_1, x_2, x_3, \ldots]^T$

These coefficients (weights) can be found using:

- Solving the model parameters analytically using closed-form equations (see here)
- An optimization algorithm such as Gradient Descent (see here)

Evaluation Metrics for Regression

Mean Absolute Error (MAE) $= \frac{1}{m}\sum_{i=1}^{m} |y^{(i)} - \hat{y}^{(i)}|$

Mean Squared Error (MSE) $= \frac{1}{m}\sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$

Root Mean Squared Error (RMSE) $= \sqrt{\frac{1}{m}\sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}$

Relative Absolute Error (RAE) $= \frac{\sum_{i=1}^{m} |y^{(i)} - \hat{y}^{(i)}|}{\sum_{i=1}^{m} |y^{(i)} - \bar{y}|}$

Relative Squared Error (RSE) $= \frac{\sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{m} \left(y^{(i)} - \bar{y}\right)^2}$

R-squared (R^2) = 1-RSE

R^2 is not error, but is a popular metric for accuracy of the model. It represents how close the data are to the fit regression line. The higher the R^2, the better the model fits your data. Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse).