Data-Driven Problem Solving in Mechanical Engineering

Model Development & Linear Regression

Masoud Masoumi

ME 364 - Spring 2022

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

March, 2022

Learning from Data

The basic premise of learning from data is the use of a set of observations to uncover an underlying process. That is a very broad premise and it is difficult to fit into a single framework.

(a) **Supervised Learning:** Training data contains explicit examples of what the correct output should be for given inputs, i.e. pairs of (input, correct output)

(b) **Unsupervised Learning:** Training data does not contain any output information. Can be viewed as the task of finding patterns and structure in input data.

(c) **Reinforcement Learning:** is concerned with the problem of finding suitable actions to take in a given situation in order to maximize a reward. Here the learning algorithm is not given examples of optimal outputs, but must instead discover them by a process of trial and error. See Gym platform.

Modeling

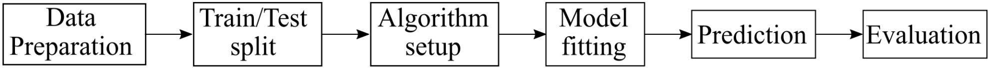
The process of encapsulating information into a tool which can forecast and make predictions.

"*All models are wrong, but some models are useful*" -George Box

To develop a model, we split the data into *training data* and *test data*, typically 80/20.

- **Training Data:** data used to fit your models or the set used for learning
- **Test Data:** data used to evaluate how good your model is.

General procedure



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:

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

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 $f(\mathrm{x}) = \omega_o + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + ...$

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

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

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) $= \frac{1}{m}\sum_{i=1}^{m}\left|y^{(i)} - \hat{y}^{(i)}\right|$

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) $= \frac{\sum_{i=1}^{m}\left|y^{(i)} - \hat{y}^{(i)}\right|}{\sum_{i=1}^{m}\left|y^{(i)} - \bar{y}\right|}$

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).