Machine Learning

Lecture 17: Regression

COURSE CODE: CSE451

2021

Course Teacher

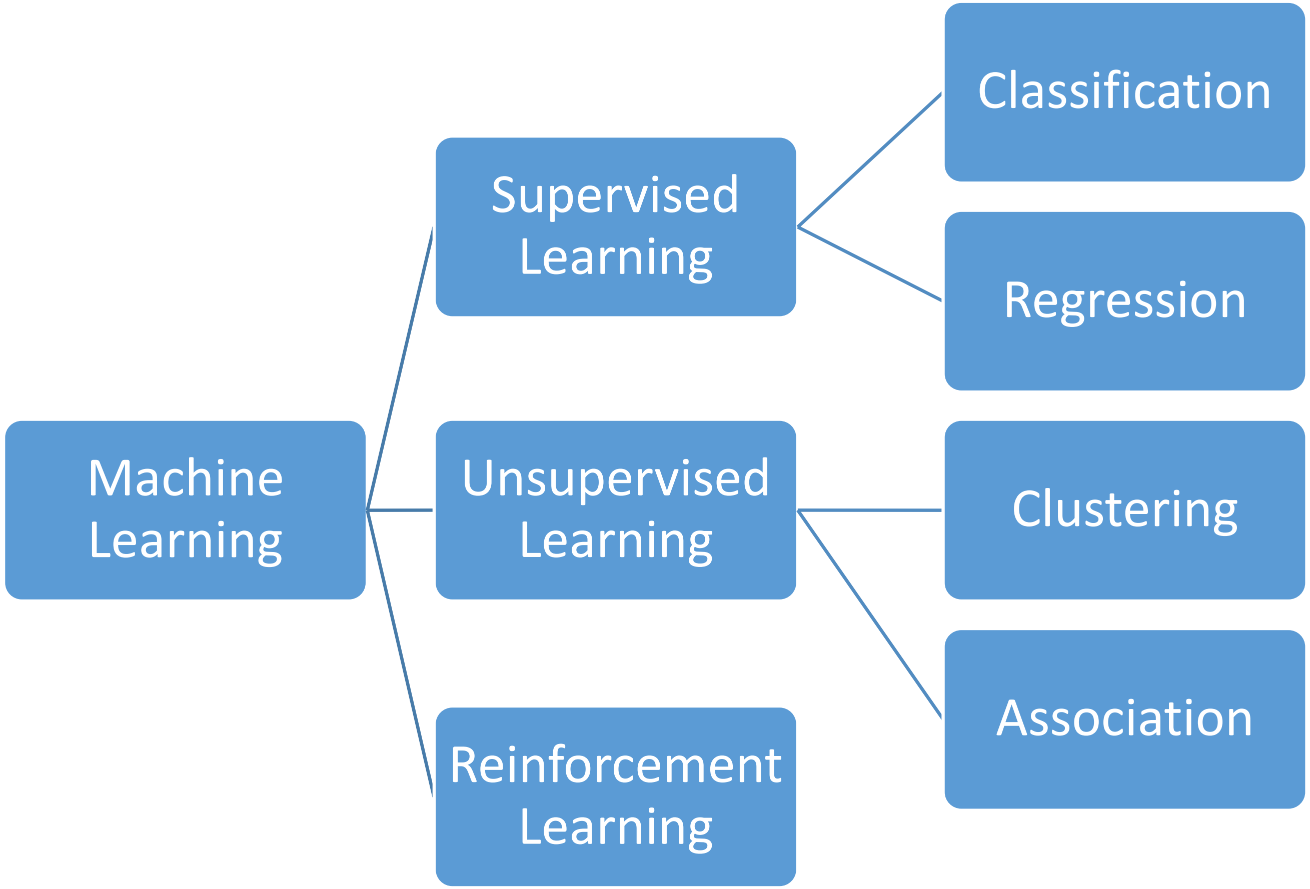
Dr. Mrinal Kanti Baowaly

Associate Professor

Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.

Email: baowaly@gmail.com

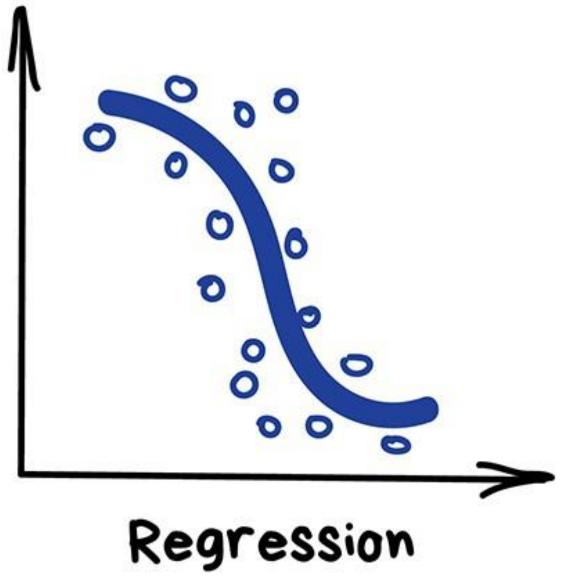
Recall: Types of Machine Learning



Source: [javaTpoint]

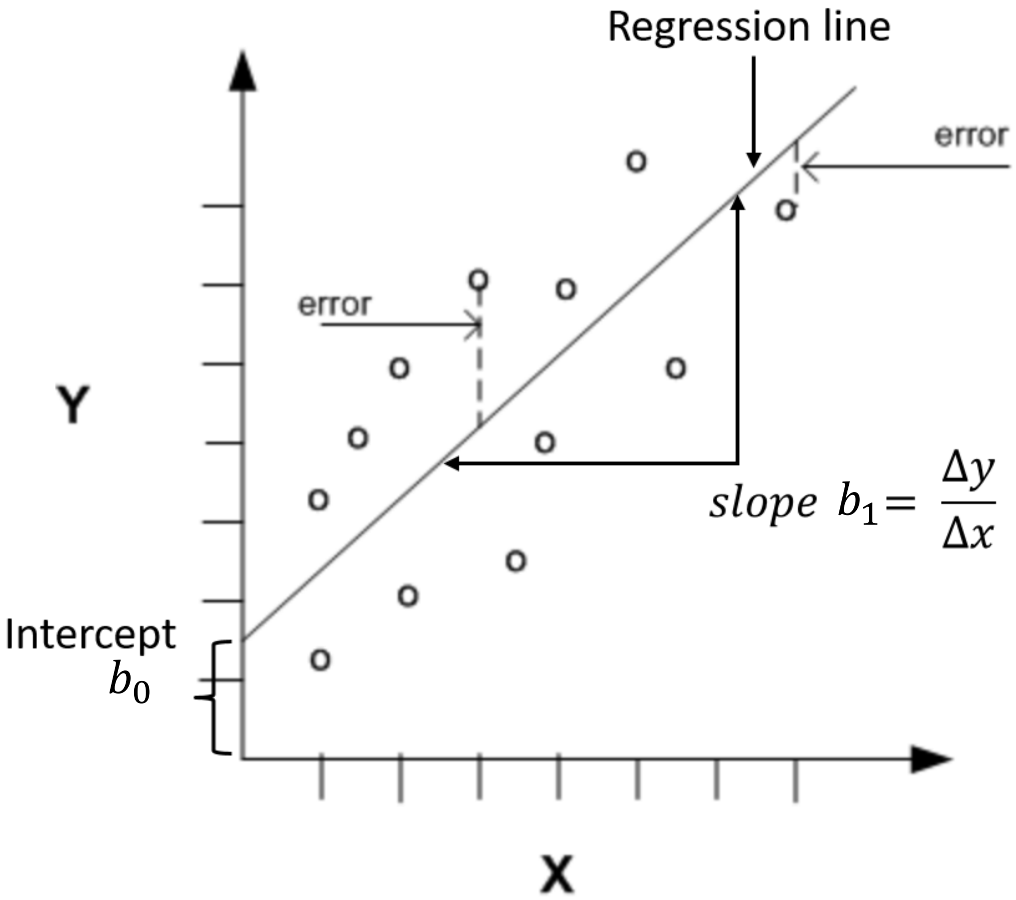
Regression

- When the output variable is a real/continuous value, such as “dollars” or “weight” or “Score”
- Predicts a single output value
- Why do we use Regression Analysis?
  - ✓Forecasting
  - ✓Demand and sales volume analysis
  - ✓Time series modelling
  - ✓Medical diagnosis etc.
- Linear Regression
- Polynomial Regression



Source: AnalyticsVidhya

How to evaluate regression model



$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – predicted value of y

$\bar{y}$ – mean value of y

Source: [AnalyticsVidhya]()

Example of calculating evaluation metrics

- Original values: -2, 1, -3, 2, 3, 5, 4, 6, 5, 6, 7
- Predicted values: -1, -1, -2, 2, 3, 4, 4, 5, 5, 7, 7
- Find:
  - ✓MAE (Mean absolute error) : 0.6363636
  - ✓MSE (Mean Squared Error): 0.8181818
  - ✓RMSE (Root Mean Squared Error): 0.904534
  - ✓R^2 (R-squared) Score: 0.9173623
- Lab Work: Use library function as well as the formulas manually to calculate these metrics

Linear Regression

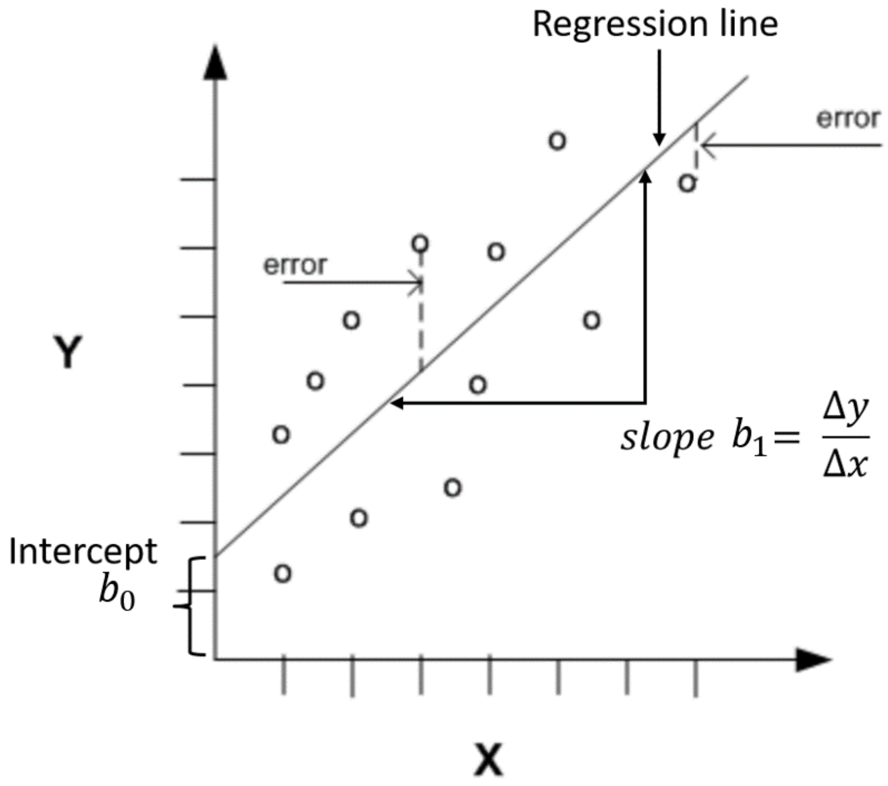
- **Linear regression** is a linear model, e.g. a model that establishes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x) using a best fit straight line (also known as regression line).
- Linear Regression is very sensitive to outliers. It can terribly affect the regression line and eventually the forecasted values.
- **Simple linear regression**: When the method has a single input variable (x)
- **Multiple linear regression:** When the method has multiple input variables (x)

Simple Linear Regression

- Simple linear regression is useful for finding relationship between two variables. One is predictor or independent variable (x) and other is response or dependent variable (y)
- The core idea is to obtain a line that best fits the data. The line can be represented by an equation:

$$y = b_0 + b_1 x$$

where y is the dependent variable, x is the independent variable, b_1 is the slope of the line and b_0 is y-intercept (constant).



Multiple Linear Regression (MLR)

- Multiple linear regression (MLR), also known simply as multiple regression, is a machine learning technique that uses several input variables to predict the outcome of a response variable.
- The goal of multiple linear regression (MLR) is to model the linear relationship between the input (independent) variables and response (dependent) variable by fitting a plane or a hyper-plane to training data. The plane can be represented by an equation as follows, given n input variables:

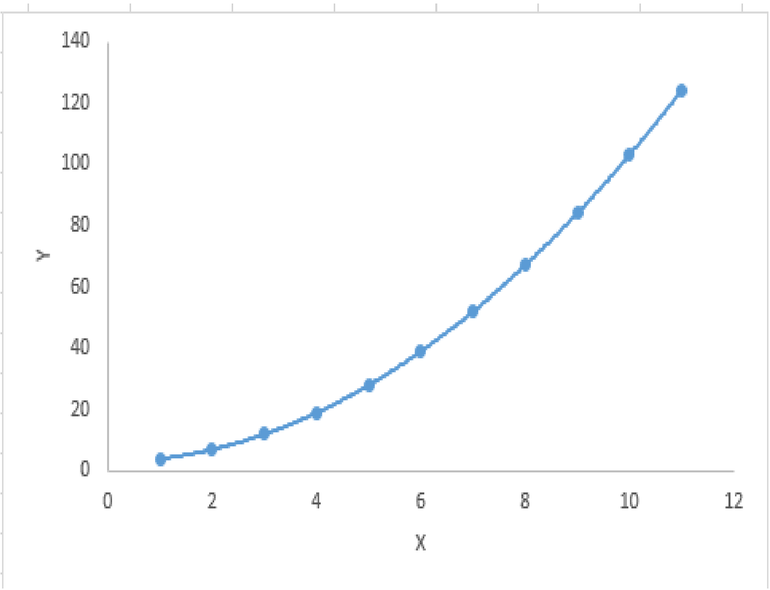
$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ \dots\dots + b_n x_{in}\ ; for\ i = 1, 2, \dots\dots n$$

where y_i is the dependent variable, x_i is the independent variable, b_i is the slope for each input variables and b_0 is y-intercept (constant).

Polynomial Regression

- A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

$$y = b_0 + b_1 x^2$$



- In regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

How to obtain best-fit regression line

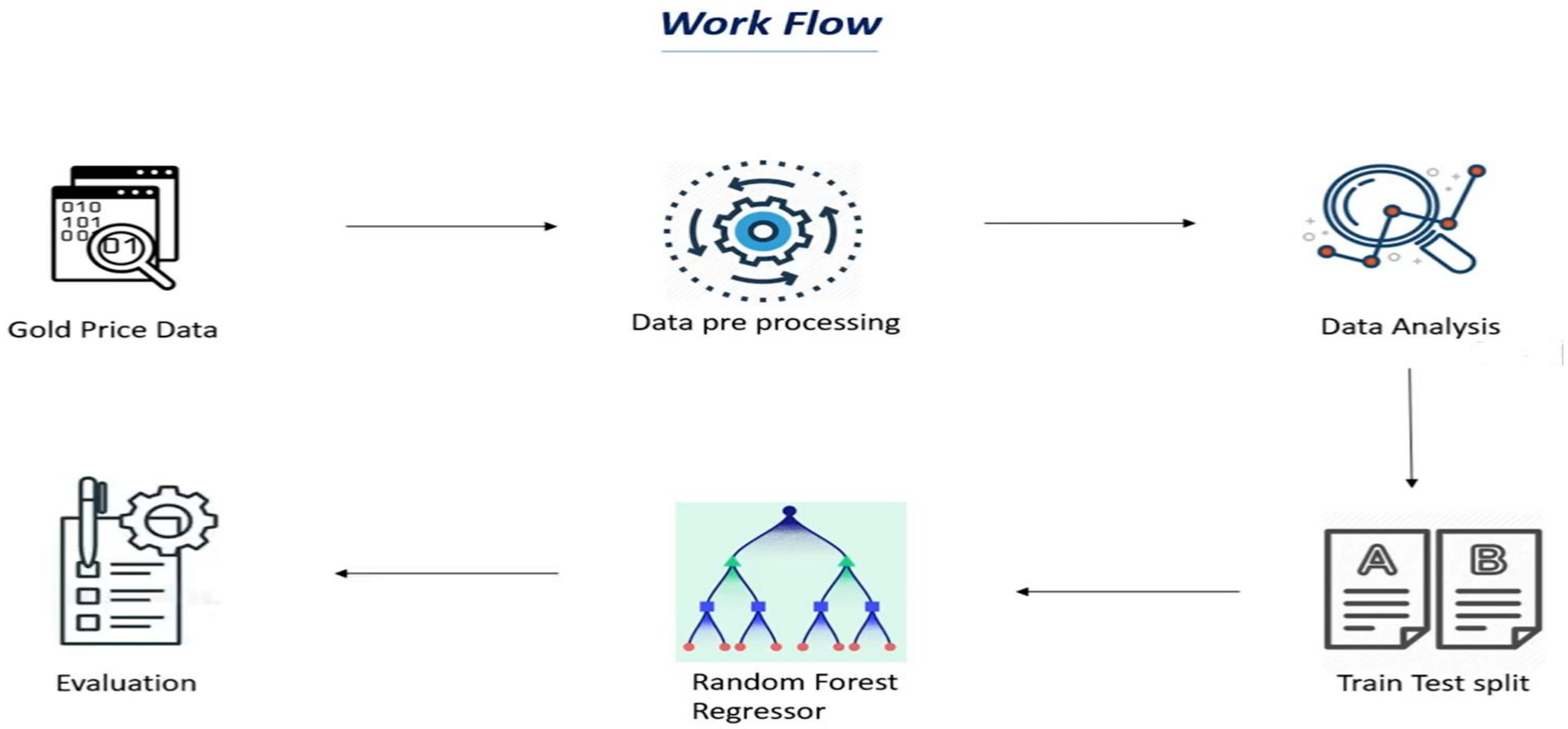
- **Cost Function (Loss Function)**: By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the coefficient values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

  Cost function of Linear Regression may be MAE (Mean absolute error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R2 (R-squared) Score etc.

- **Gradient Descent:** This is a process of optimizing the values of the coefficients by iteratively minimizing the cost function of your model. The idea is to start with random values for each coefficient and then iteratively update the values until reaching minimum cost (error).

An Example Project on Regression

- Building A Gold Price Prediction Model Using Machine Learning



[Dataset Link]
Prediction: [AV], [Kaggle]

Linear Regression vs Logistic Regression

- Linear Regression is used for solving Regression problem. It is used to predict the continuous dependent variable using a given set of independent variables.
- Logistic regression is used for solving Classification problems. It is used to predict the categorical dependent variable using a given set of independent variables.

Detail: JavaTpoint

Study Materials of Regression

7 Regression Techniques you should know!

Know The Best Evaluation Metrics for Your Regression Model

Linear Regression for Machine Learning

Linear Regression in Machine Learning, Simple Linear Regression, Multiple Linear Regression, Polynomial Regression