TA 2

Arun Kumar Rajasekaran

What's ML?

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy.

How does ML work?

1. A Decision Process: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labeled or unlabeled, your algorithm will produce an estimate about a pattern in the data.

2. An Error Function: An error function evaluates the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.

3. A Model Optimization Process: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this iterative “evaluate and optimize” process, updating weights autonomously until a threshold of accuracy has been met.

4 types of ML algorithms

- Supervised
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Supervised Learning

Supervised learning is a machine learning approach where algorithms learn from labeled data. The algorithm receives input data and corresponding correct output labels in this process. The objective is to train the algorithm to predict accurate labels for new, unseen data. Examples of supervised learning algorithms include:

- Decision Trees
- Support Vector Machines
- Random Forests
- Naive Bayes

These algorithms can be used for classification, regression, and time series forecasting tasks. Supervised learning is widely used in various domains, including healthcare, finance, marketing, and image recognition, to make predictions and gain valuable insights from data.

Unsupervised Learning

In this machine learning approach, algorithms analyze unlabeled data without predefined output labels. The objective is to discover patterns, relationships, or structures within the data. Unlike supervised learning, unsupervised learning algorithms work independently to uncover hidden insights and group similar data points together. Common unsupervised learning techniques include clustering algorithms like:

- K-means
- Hierarchical clustering
- Dimensionality Reduction Methods like PCA and t-SNE

Semi supervised Learning

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of not having enough labeled data for a supervised learning algorithm. It also helps if it’s too costly to label enough data.

Reinforcement Learning

Reinforcement machine learning is a machine learning model that is similar to supervised learning, but the algorithm isn’t trained using sample data. This model learns as it goes by using trial and error. A sequence of successful outcomes will be reinforced to develop the best recommendation or policy for a given problem.

Linear Regression

In simple words

Simple linear regression is used to estimate the relationship between two quantitative variables.

- How strong the relationship is between two variables (e.g., the relationship between rainfall and soil erosion).
- The value of the dependent variable at a certain value of the independent variable (e.g., the amount of soil erosion at a certain level of rainfall).

Linear Regression

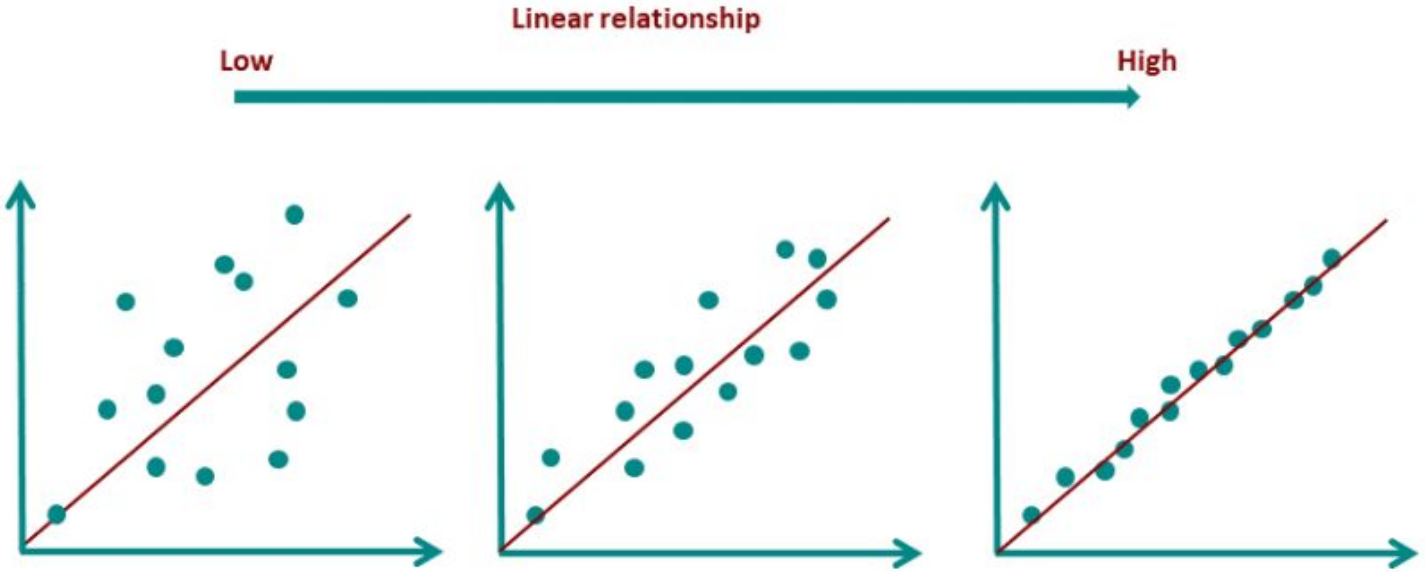
The regression line can be described by the following equation:

$$\hat{y} = b \cdot x + a$$

- $\hat{y}$: Estimated dependent variable
- b: Slope
- x: Independent variable
- a: y intercept

Definition of "Regression coefficients":

- ***a*** : point of intersection with the *y*-axis
- ***b*** : gradient of the straight line



Assumptions of Linear Regression

Generic ‘Least square method’ (ref. Next slide)

Step 1: *Calculate the slope ‘m’ by using the following formula:*

$$m = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$b = \frac{\sum(x - \bar{x}) * (y - \bar{y})}{\sum(x - \bar{x})^2}$$

Step 2: *Compute the y-intercept (the value of y at the point where the line crosses the y-axis):*

$$c = y - mx$$

Step 3: *Substitute the values in the final equation:*

$$y = mx + c$$

Assumptions of Linear Regression

https://www.technologynetworks.com/informatics/articles/calculating-a-least-squares-regression-line-equation-example-explanation-310265

Assumptions of Linear Regression

- Linearity: There must be a linear relationship between the dependent and independent variables.
- Homoscedasticity: The residuals must have a constant variance.
- Normality: Normally distributed error
- No multicollinearity: No high correlation between the independent variables
- No auto-correlation: The error component should have no auto-correlation

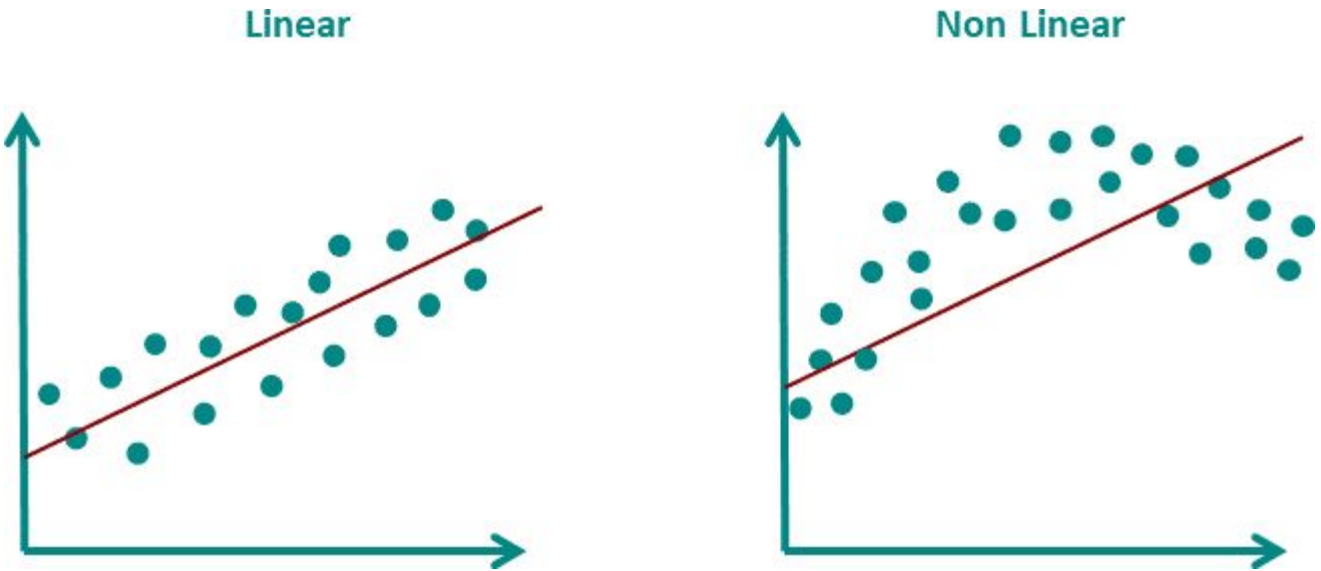
A bit more on the assumptions

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. Normality: The data follows a normal distribution.

Linear regression makes one additional assumption:

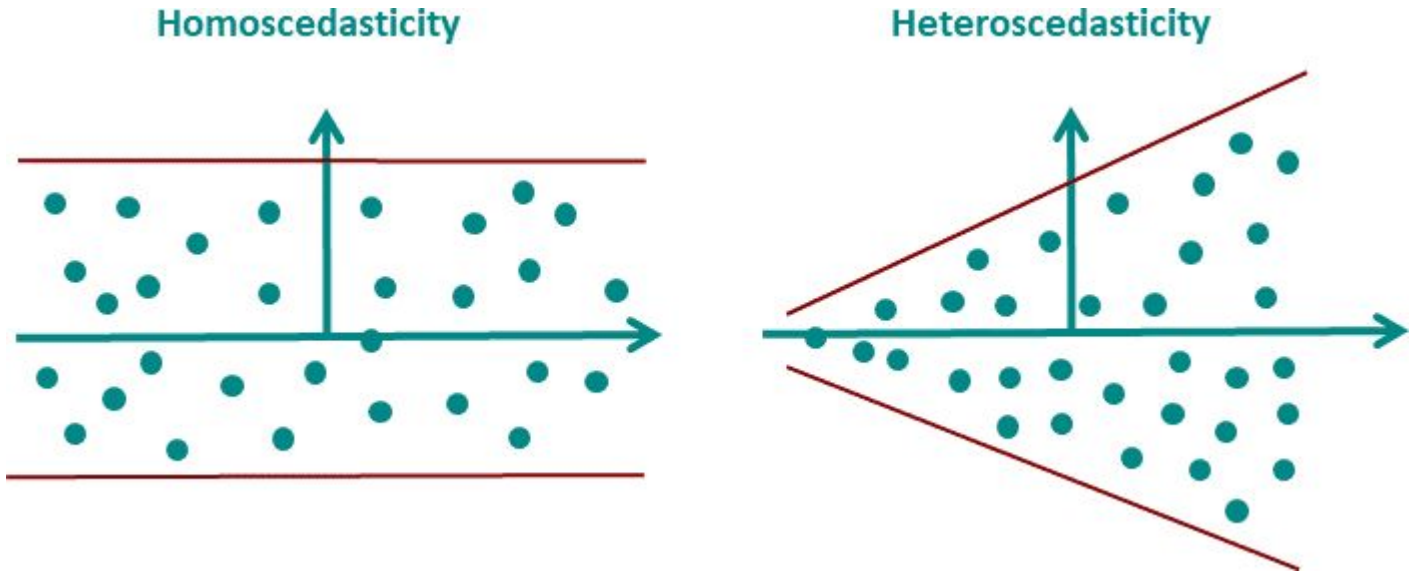
4. The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

1. Linearity



In linear regression, a straight line is drawn through the data. This straight line should represent all points as good as possible. If the points are distributed in a non-linear way, the straight line cannot fulfill this task.

2. Homoscedasticity



Since in practice the regression model never exactly predicts the dependent variable, there is always an error. This very error must have a constant variance over the predicted range.

2. Homoscedasticity

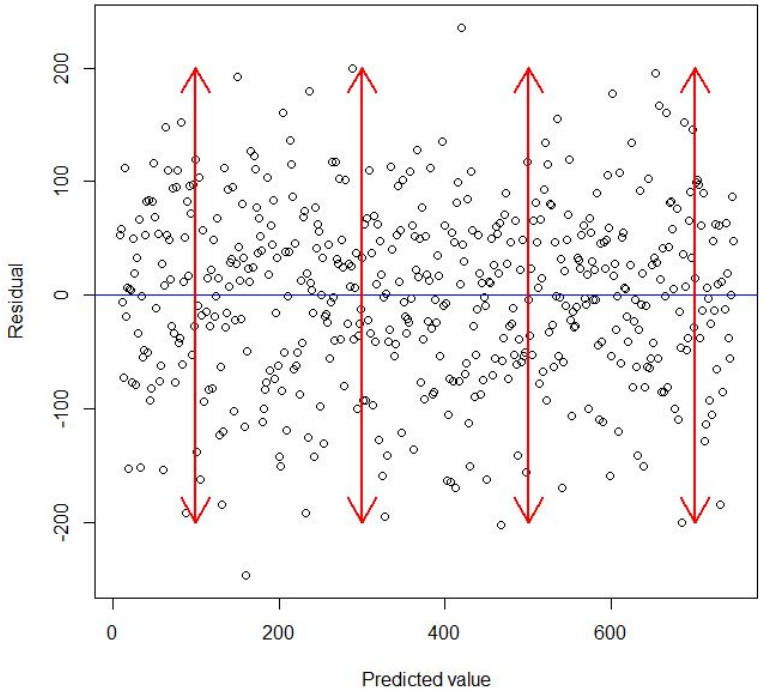
Advanced analysis include,

***F*-Test**

Modified Levene Test

Breusch-Pagan Test

Bartlett's Test



2. Homoscedasticity

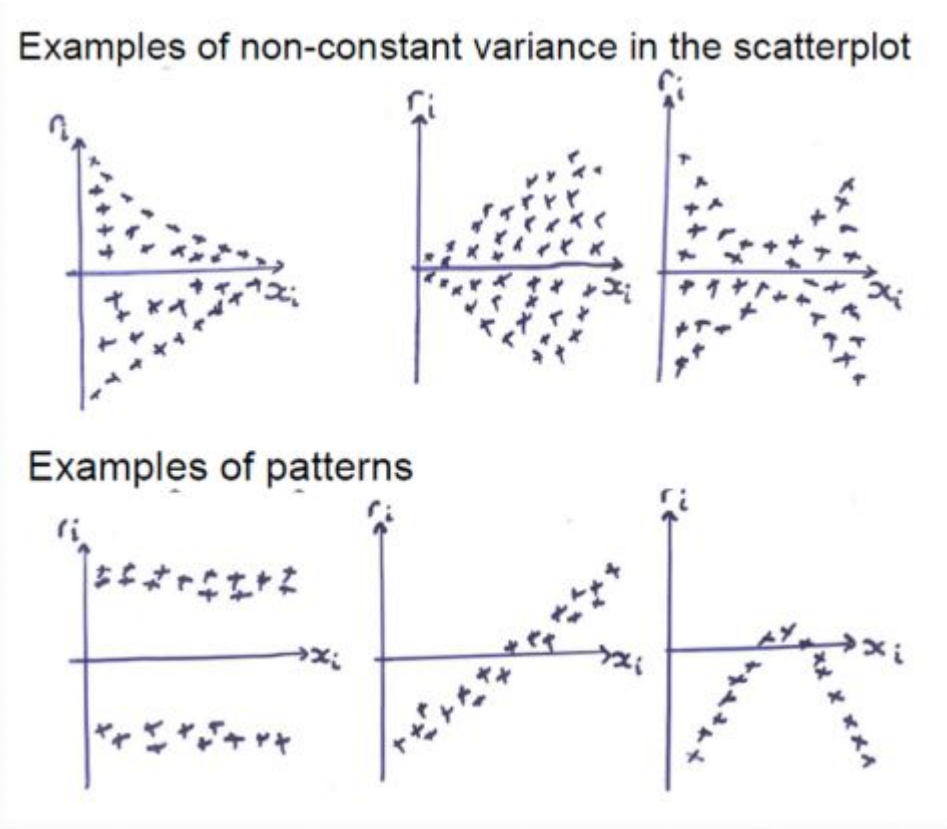
Advanced analysis include,

***F*-Test**

Modified Levene Test

Breusch-Pagan Test

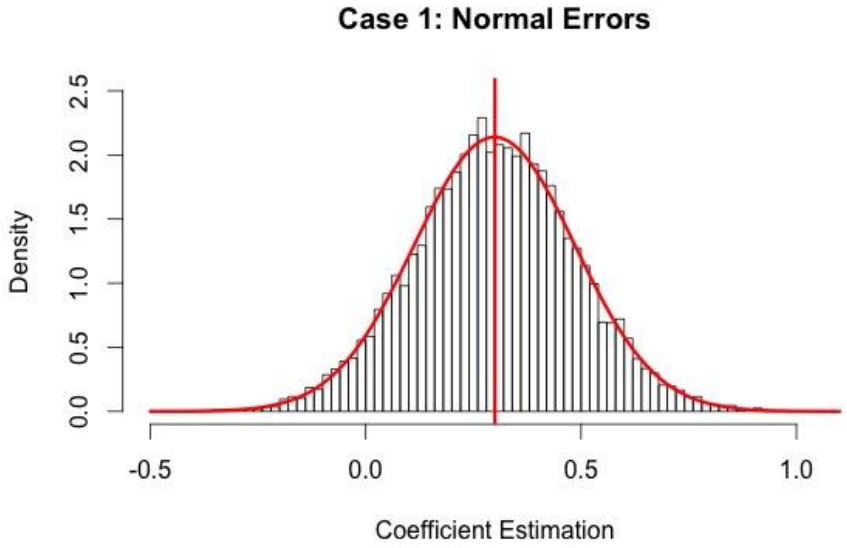
Bartlett's Test

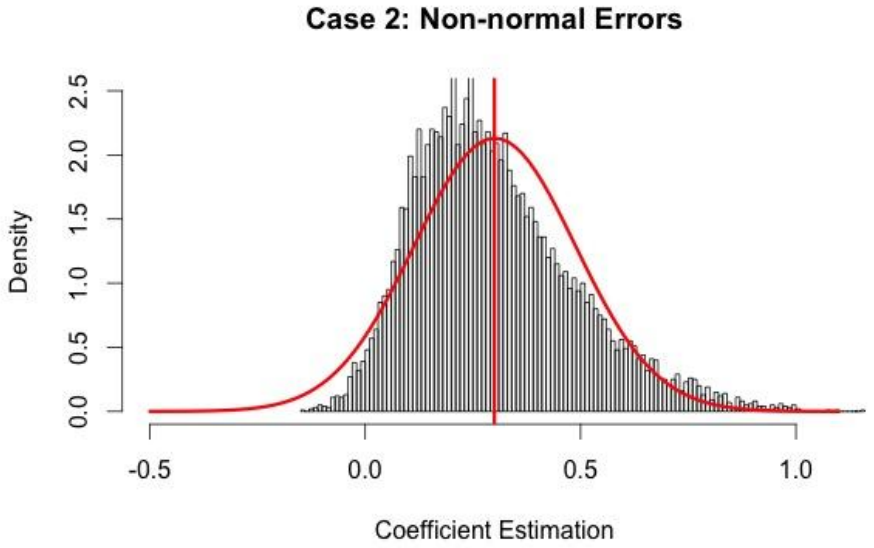


2. Homoscedasticity (how to fix?)

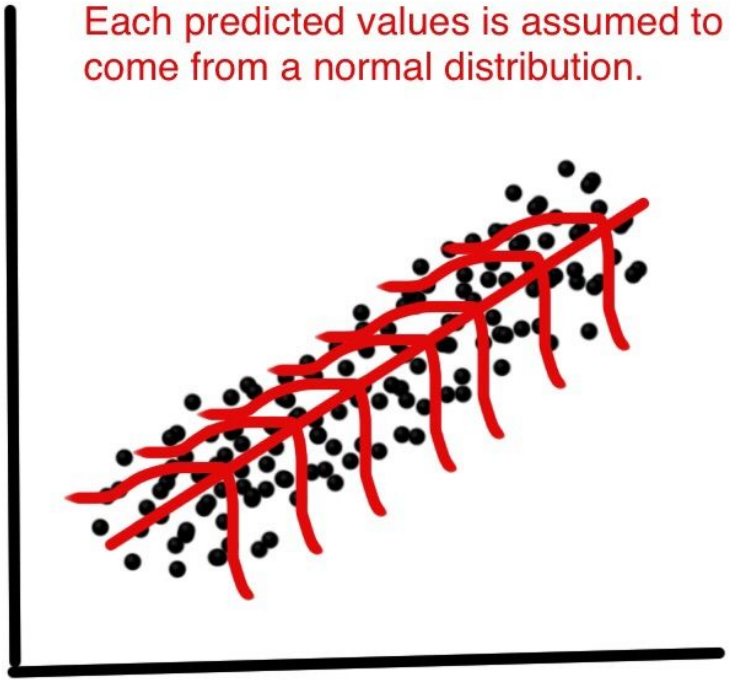
- **Variance stabilizing transformation:** A transformation of the outcome used to correct non-constant variance is called a “variance stabilizing transformation.” common transformations are the natural logarithm, square root, inverse, and Box-Cox
- Advanced methods such as weighted or generalized least squares can be used to handle non-constant variance.
- Non-constant variance may co-occur with non-linearity and/or non-normality.

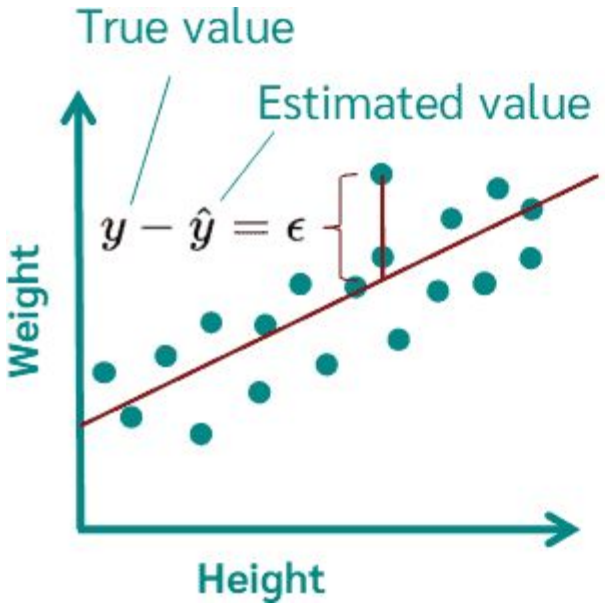
3. Normal distribution of error





3. Normal distribution of error





Error epsilon

$$y = b \cdot x + a + \epsilon$$

Multiple LR vs Multivariate

Simply…

“Regression analysis results in a formula of the form Y=a+bX. A multiple regression has more than one X in one formula. A multivariate regression has more than one Y, but in different formulae. And a multivariate multiple regression has multiple X’s to predict multiple Y’s with each Y in a different formula, usually based on the same data.”

A bit more, equations…

Simple regression pertains to one dependent variable (y) and one independent variable (x): $y=f(x)$

Multiple regression (aka multivariable regression) pertains to one dependent variable and multiple independent variables: $y=f(x1,x2,...,xn)$

Multivariate regression pertains to multiple dependent variables and multiple independent variables: $y1,y2,...,ym=f(x1,x2,...,xn)$

.