Machine Learning _101

A complete crash course

COSC

Table of contents

01

Machine Learning?

A Beginners Guide

02

Types of Learnings

The power of a trained model

03

Supervised Learning

Classification vs Regression

04

Unsupervised Learning

Have you ever heard of clustering?

1943

The **year** where the idea of Machine Learning was conceived from a research paper on mathematical modeling of neural networks by logician Walter Pitts

So What Is Machine Learning?

- Writing software is the bottleneck: Traditional software development can be time-consuming and labor-intensive, requiring extensive coding and testing.
- Getting computers to program themselves: The future of automation lies in leveraging machine learning and artificial intelligence (AI) to automate the process of creating software.
- Let the data do the work instead!: By utilizing data-driven approaches, we can train models to automatically generate code, making software development faster, more efficient, and less error-prone.

Machine Learning

Also known as predictive analysis

It is used to analyze past trends in order to predict likelihood of future outcomes.

Used in -

- Voice recognitions
- Email Filtering
- Autopilot systems etc..

Sample Applications

- Web search
- Computational biology
- Finance
- E-commerce
- Space exploration
- Robotics
- Information extraction
- Social networks
- Debugging

Data?

Machine is learning ?

A) When computer is able to figure out a relationships between input variables.

Train Model (Model that has learned right set of instructions for a given task)

Train data

Test data

Labeled data, unlabeled data

Features / input data , Target

Types of Learning

- Supervised (inductive) learning
    - Training data includes desired outputs
- Unsupervised learning
    - Training data does not include desired outputs
- Semi-supervised learning
    - Training data includes a few desired outputs
- Reinforcement learning
    - Rewards from sequence of actions

Supervised Learning

It is a process of training a predictive model.

Predictive models are used to assign labels to unlabeled data based on patterns learned from it.

Used in-

- Image recognition
- Text prediction
- Spam filtering etc..

Data Types...

Categorical data

- Attribute that holds data stored in discrete form.
- Ex - Customer Name , Grade etc

Continuous data

- Attributes that hold data stored in the form of integer or real numbers.
- Has infinite number of possible values between its lower & upper bounds

Numeric scale ranging from 1 to 5 is categorical ?

Types..

Regression

- Regression algorithms are used to determine continuous values such as price, income, age, etc.
- Regression Algorithms - Linear Regression , random forest etc..

Classification

- Classification algorithms are used to forecast or classify the distinct values such as Real or False, Male or Female, Spam or Not Spam, etc.
- Ex- Logistic Regression , K nearest neighbour etc..

Unsupervised Learning

The machine evaluates input data and identifies any hidden patterns (or) relationships that exists in unlabelled data.

Used-

- Movie recommendation systems
- Customer Segmentation process etc..

Types..

Clustering

- Method of grouping the objects into clusters so that objects with the most similarities remain in groups.

Association

- An association rule is an unsupervised learning method used to find the relationships between variables in the large database.
- It determines the set of items that occurs together in the dataset.

Semi-Supervised Learning

Semi-supervised learning is a learning problem that involves a small number of labeled examples and a large number of unlabeled examples.

Ex-

- Web content classification
- Text document classification
- Speech recognition

Reinforcement Learning

Agent

Environment

Agent <—--Action & feedback—--> Environment

Used in-

- Computer Games
- Trading

Steps in Machine learning process

- Data collection
- Data Exploration
- Data preparation
- Modelling
- Evaluation
- Actionable insight

Data Collection

Identify and gather information for required model

- Unsupervised learning (unlabelled data)
- Supervised learning (historical labelled data)

Ex- Forming cricket team

Data Exploration

Describing, visualizing and analyzing data.

Includes-

- Number of rows & columns in data
- Type of values stored in columns
- Any missing values/duplicates/outliners

Ex - Checking team members eligibility

Data Preparation

Modify data as per our model

- Includes-
- Dealing with missing data , outliers etc..
- Modifying/transforming the structure of data

Ex- Replacing team members

Handling missing values

1. Simply remove all instances with features that have a missing value.
2. Use indicated values such as NA , unknown or negative one (-1) to represent missing values.
3. Imputation - Use of a systematic approach to fill in missing values by using most probable substitute values (Median imputation , Mean imputation etc..)

Modelling & Evaluation

Modelling

- Choosing/finding right ML model that works for given data
- Ex- Playing test matches

Evaluation

- Access how well machine learning approach worked
- Ex- Evaluating player’s performance

Actionable Insight

Identify what to do based on results of the machine learning approach you choose

Ex-

- Unsupervised learning - what to do with patterns?
- Supervised learning - Deciding whether to deploy or not?

What We’ll Cover

- **Supervised learning**
    - Regression: Linear Regression
    - Classification: Logistic and Decision Trees

- **Unsupervised learning**
    - Clustering

Bias & Variance

Bias: Error rate of the training data

- High bias
- Low bias

- **Variance:** The difference between the error rate of training data and testing data is called variance.
- High variance
- Low variance

Overfitting

Overfitting: A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance.

Reasons for Overfitting are as follows:

- High variance and low bias
- The model is too complex
- The size of the training data

Underfitting

Underfitting: A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data.

Reasons for Underfitting:

- High bias and low variance
- The size of the training dataset used is not enough.
- The model is too simple.
- Training data is not cleaned and also contains noise in it.

Best fit/Good fit

Good Fit/best fit

- Ideally, the case when the model makes the predictions with 0 error, is said to have a *good fit* on the data. This situation is achievable at a spot between overfitting and underfitting.

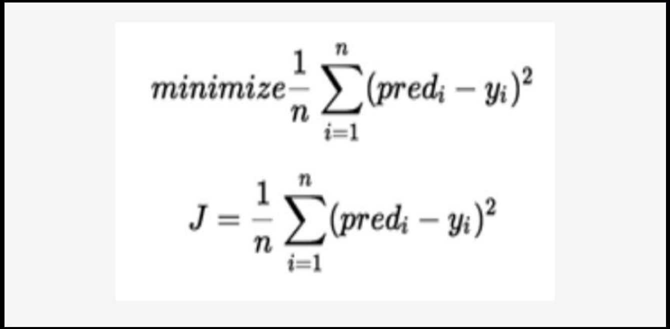
Supervised Learning_

"It uses machine learning algorithms to analyze and cluster labeled instances of data, say $\langle x_i, y \rangle$ "

Terminologies

Cost Function

- The output which is obtained or predicted by an algorithm is referred to as yˆ
- The difference between the actual and predicted values is the error, i.e., y - yˆy^. Different values of y- yˆy^ is loss function.
- The average summation of all loss function values is called the cost function.
- The machine learning algorithm tries to obtain the minimum value of the cost function.



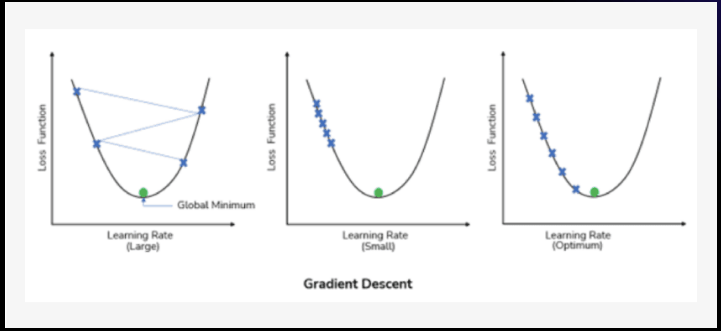
$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Continuation...

Gradient Descent

- It is a popular optimization approach employed in training machine learning models by reducing errors between actual and predicted outcomes.
- Optimization in machine learning is the task of minimizing the cost function parameterized by the model's parameters.
- The primary goal of gradient descent is to minimize the convex function by parameter iteration.

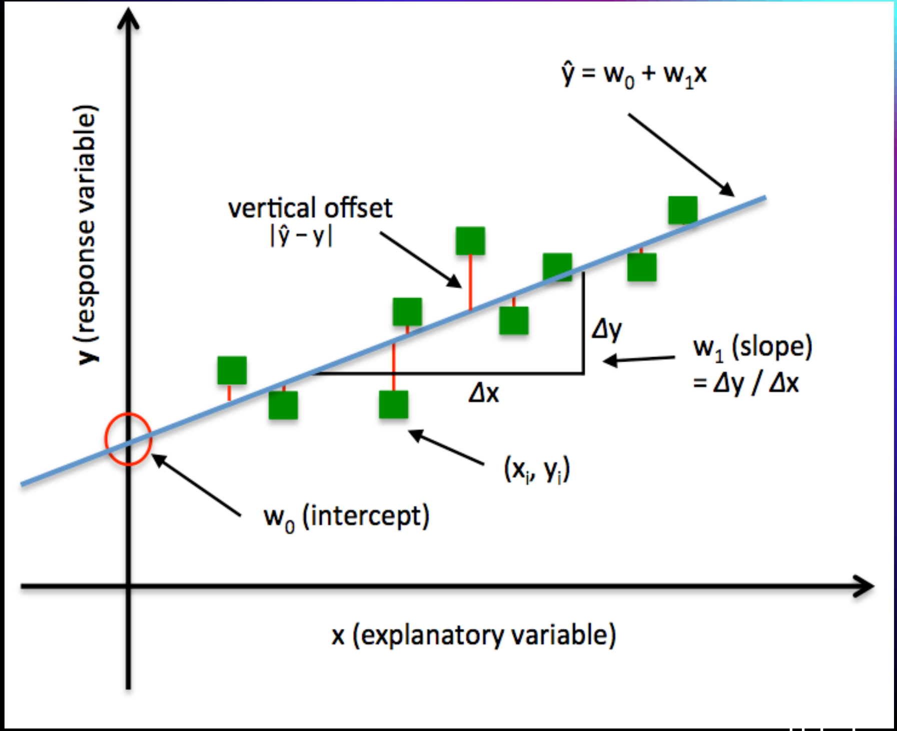


Linear Regression

- This is the base model for all statistical machine learning
- **x** is a one feature data variable
- y is the value we are trying to predict
- **The regression model is**

$$y = w_0 + w_1 x + \varepsilon$$

- Two parameters to estimate – the slope of the line w1 and the y-intercept w0
- ε is the unexplained, random, or error component.



Regression Types..

Simple Linear Regression

- A simple straight-line equation involving slope (dy/dx) and intercept (an integer/continuous value) is utilized in simple Linear Regression.
- Here a simple form is: y=mx+c

  where y denotes the output x is the independent variable, and c is the intercept when x=0.
- With this equation, the algorithm trains the model of machine learning and gives the most accurate output

Continuation..

Multiple Linear Regression

- When a number of independent variables more than one, the governing linear equation.
- It takes a different form like: y= c+m1x1+m2x2… mnxn where represents the coefficient responsible for impact of different independent variables x1, x2 etc.

Non-Linear Regression

- When the best fitting line is not a straight line but a curve, it is referred to as Non-Linear Regression.

Logistic Regression

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.
- The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes.
- Logistics regression uses the sigmoid function to return the probability of a label.

Sigmoid Function

- Sigmoid Function is a mathematical function used to map the predicted values to probabilities. The function has the ability to map any real value into another value within a range of 0 and 1.

Types..

- **Binomial**: Where the target variable can have only two possible types.

  Ex- Predicting a mail as spam or not.

- **Multinomial**: Where the target variable have three or more possible types, which may not have any quantitative significance.

  Ex- Predicting disease.

- **Ordinal**: Where the target variables have ordered categories.

  Ex- Web Series ratings from 1 to 5.

Decision Trees

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
- It is a tree-structured classifier, where

  Internal nodes (represents the features of a dataset)

  Branches (represents the decision rules)

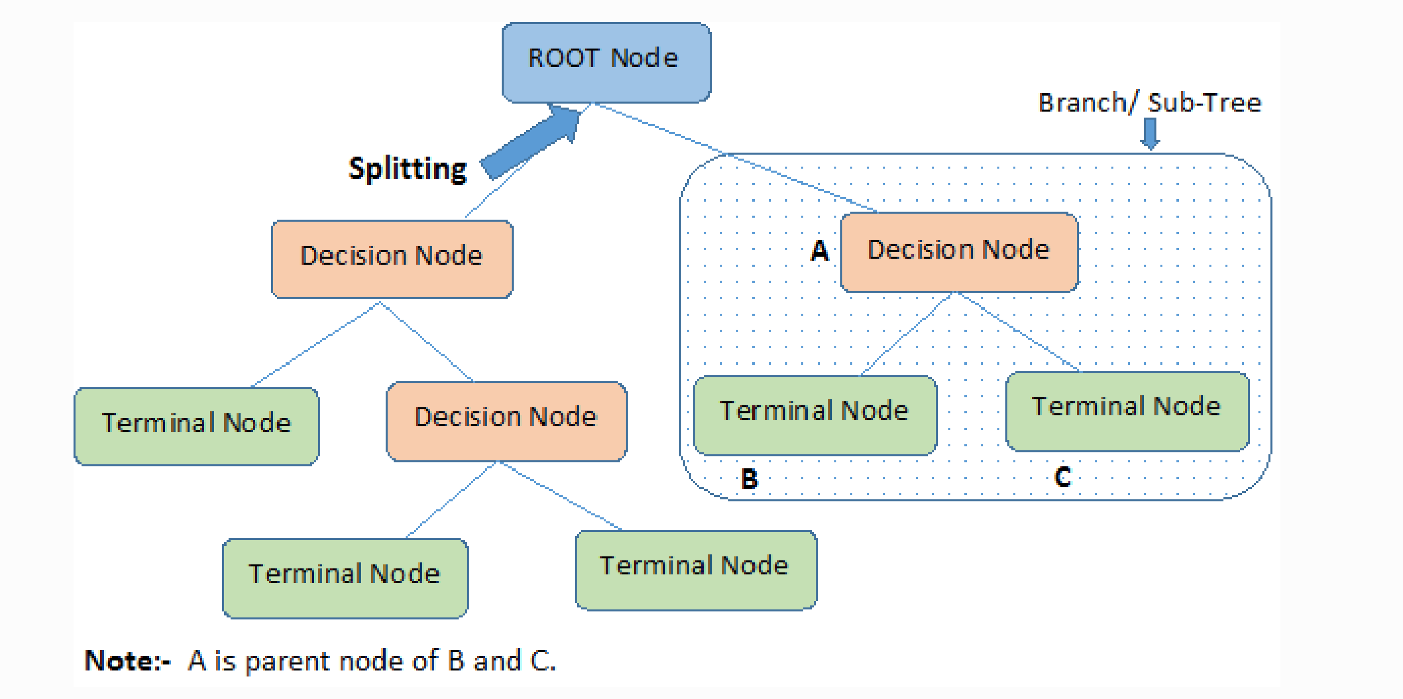
  Leaf node (represents the outcome)
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node.

  Decision nodes are used to make any decision and have multiple branches.

  Leaf nodes are the output of those decisions and do not contain any further branches.

Important Terminologies

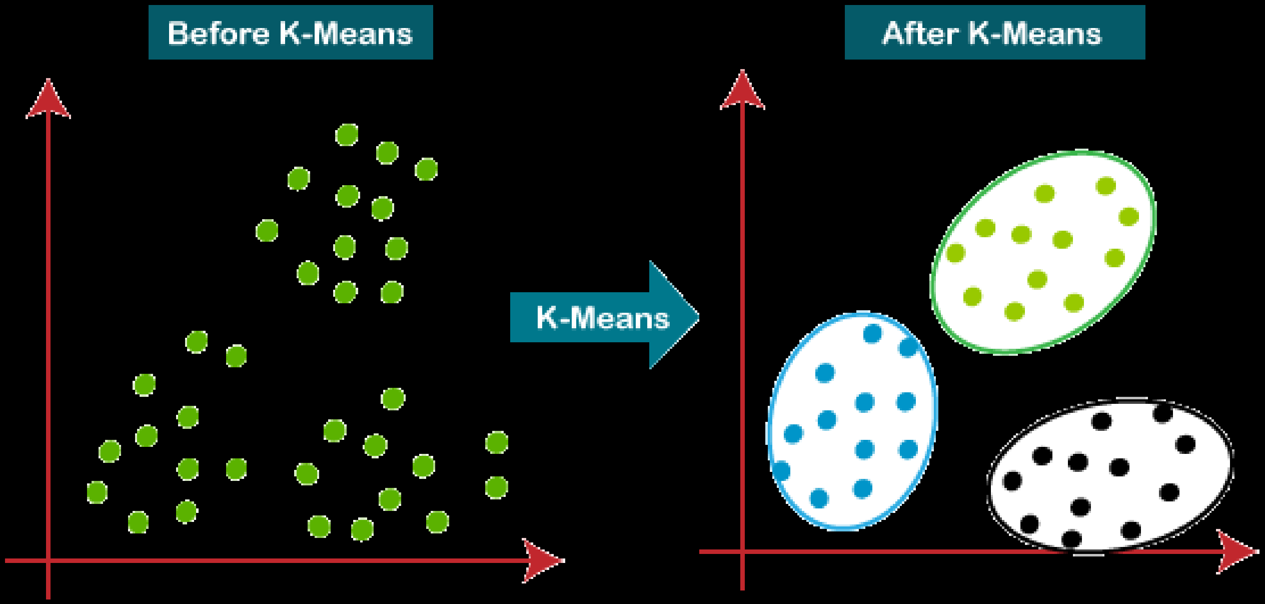
- **Root Node:** It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.
- **Leaf / Terminal Node:** Nodes do not split is called Leaf or Terminal node.
- **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
- **Branch / Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.
- **Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.



Note:- A is parent node of B and C.

K-Means Clustering

- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.
- K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid to assign it to a cluster. The goal is to identify the K number of groups in the dataset.
- The term ‘K’ is a number. You need to tell the system how many clusters you need to create. For example, K = 2 refers to two clusters. There is a way of finding out what is the best or optimum value of K for a given data.


Before K-Means
K-Means
After K-Means

THANK YOU!