Chapter -2 Fundamentals of Machine Learning

Types of ML Algorithms

In general, machine learning algorithms can be classified into three types.

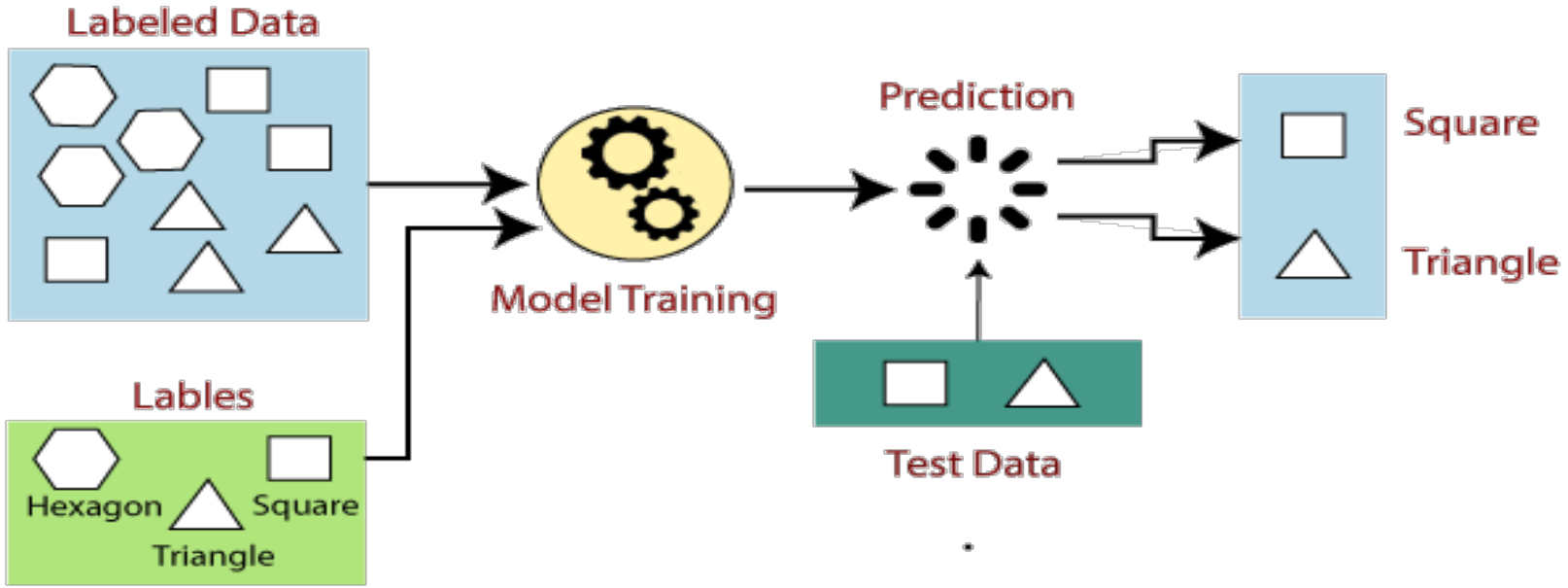
- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised Learning

- Supervised learning involves **training an algorithm** on a labelled,classified or categorized dataset, where input data is paired with corresponding output labels.
- The goal is to learn a mapping from input to output based on provided labelled examples.
- The aim is to map input **variable x to output variable y**
- **y=f(x)**

How Supervised Learning works

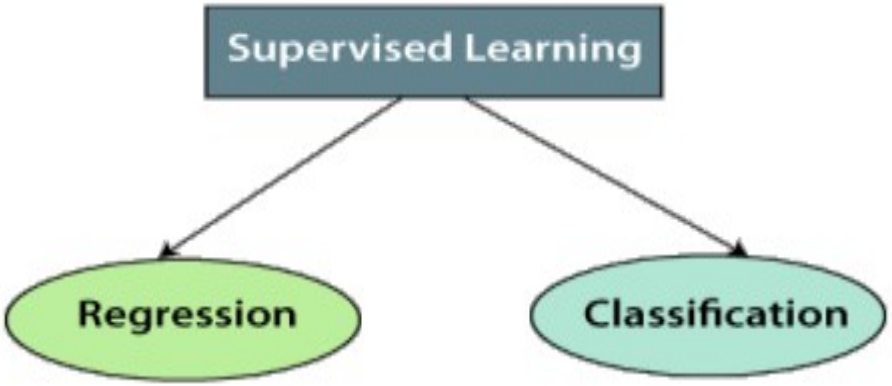
- models are **trained using** labelled dataset, where the model learns about each type of data.
- Once the training process is completed, the model is **tested on the basis** of test data (a subset of the dataset), and then it predicts the output.



Steps Invloved In Supervised Learning

- First determine the **type of training dataset**
- **Collect/Gather** the labeled training data.
- **Split the dataset** into training dataset, test dataset, and validation dataset.
- Determine **the input features** of the training dataset
- **Determine the suitable algorithm for the model**, such as svm, decision tree, etc.
- **Execute the algorithm** on the training dataset.
- **Evaluate the accuracy** of the model by providing the test set.

Types of Supervised Learning



Regression Algorithms

Regression algorithms are used if there is a **relationship** between the **input variable and the output variable.**

- It is used for the prediction of **continuous variables**, such as weather forecasting, Market Trends, etc.
  - **Linear Regression**
  - **Non-Linear Regression**
  - **Regression Trees**
  - **Bayesian Linear Regression**
  - **Polynomial Regression**

Supervised Learning classification Algor.

Classification algorithms are used when t**he output variable is categorical**

- there are **classes** such as Yes-No, Male-Female, True-false, Low-Middle-High etc. e.g.Spam Filtering,

- **Examples of Classification algorithms**

  - **Decision Trees**
  - **Random Forest**
  - **Logistic Regression**
  - **Support Vector Machines**

Adv. Of Supervised Learning

- With the help of supervised learning, the model can predict **the output on the basis of** prior experiences.
- In supervised learning, we can have an **exact idea about the classes** of objects.
- Supervised learning model **helps us to solve various real-world problems** such as fraud detection, spam filtering, etc.

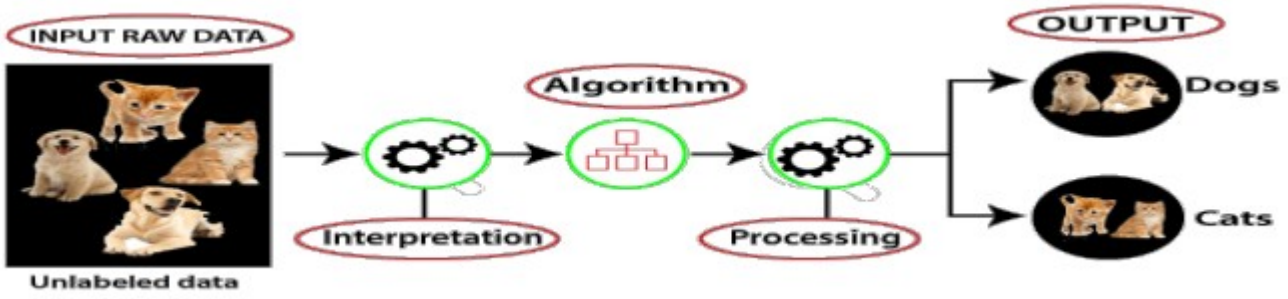
Unsupervised Learning

- Unsupervised learning involves training a model on **not labelled, classified, or categorized**, and the algorithm needs to act on that data without any supervision,
- The algorithm aims to **explore the inherent structure** in the data.
- models itself **find the hidden patterns and insights** from the given data.

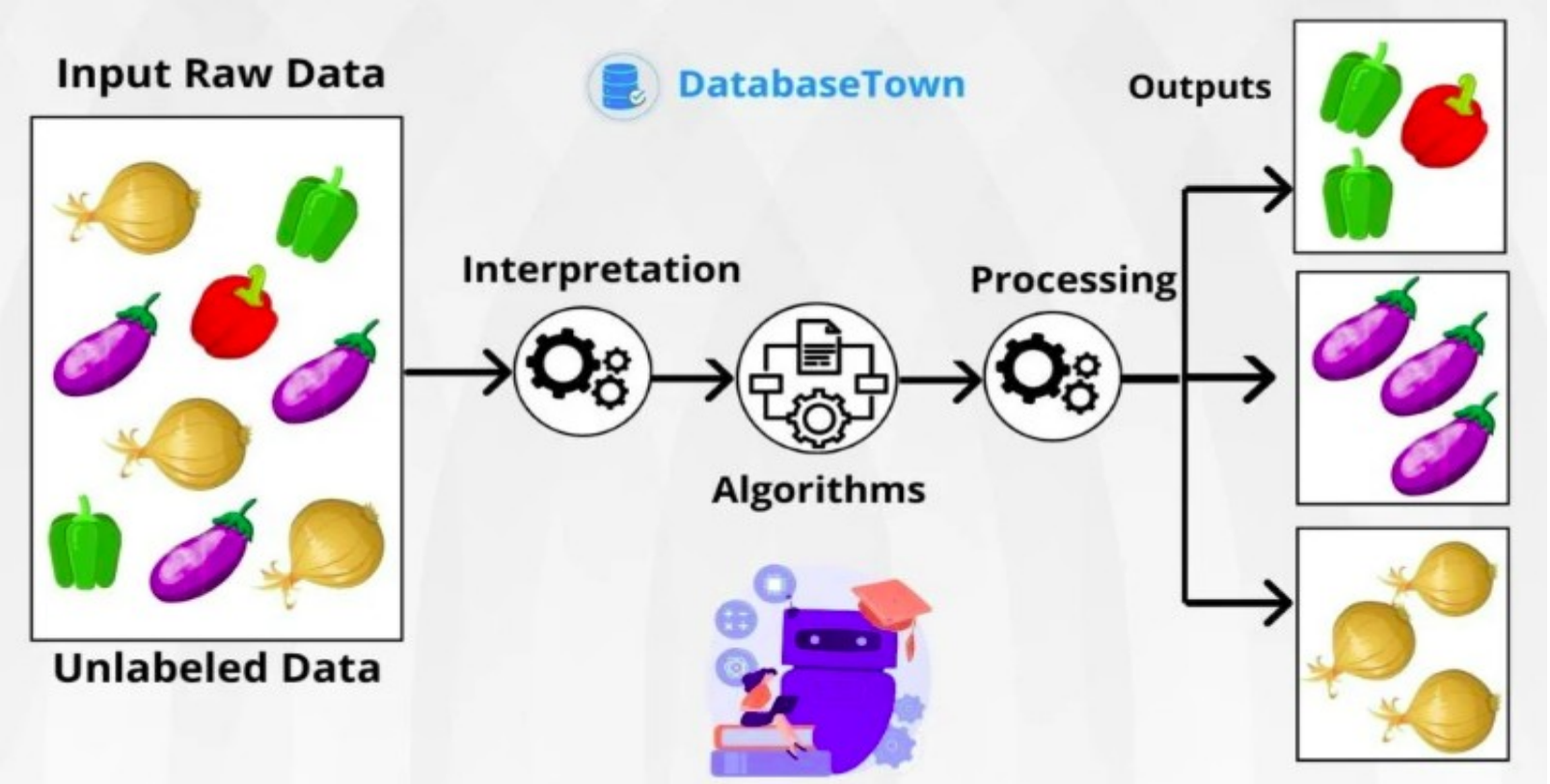
Unsupervised Learning

- Suppose the unsupervised learning algorithm is given an input dataset containing **images of different types** of cats and dogs

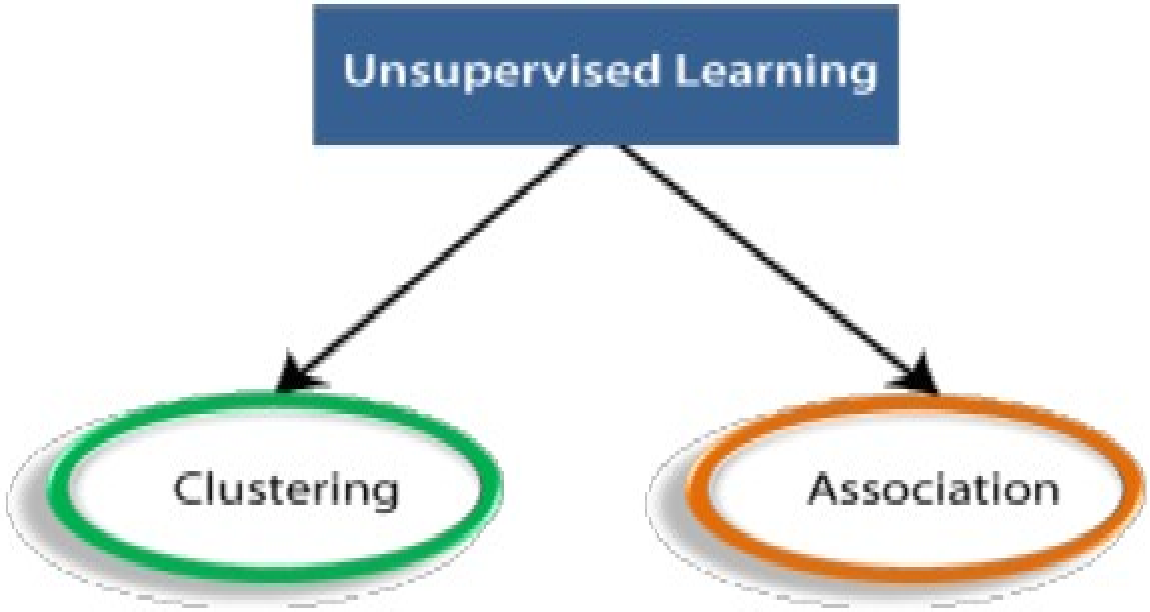
- The Algorithm **does not have any idea about the features** of the dataset. Identifies image features by its own and group them



Unsupervised Learning



Unsupervised Learning



Unsupervised ML Problem Types

A. Clustering:

Clustering is a method of **grouping the objects into clusters** such that objects **with most similarities** remains into a group

Ex: Market segmentation and anomaly detection(find unusual network traffic)

B. Association :

An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database

find associations and patterns in the data customer who buy x will also buy y

ex : dimensionality reduction

Unsupervised ML Algorithms

- K-Means Clustering
- Hierarchal Clustering
- Anomaly Detection
- Neural Networks
- Principle Component Analysis
- Apriori Algorithm
- Singular Value Decomposition

Advantages of Unsupervised Learning

- Unsupervised learning is used for **more complex tasks** as compared to supervised learning
- Unsupervised learning is preferable as it is easy to **get unlabeled data** in comparison to labeled data.

Disadvantage

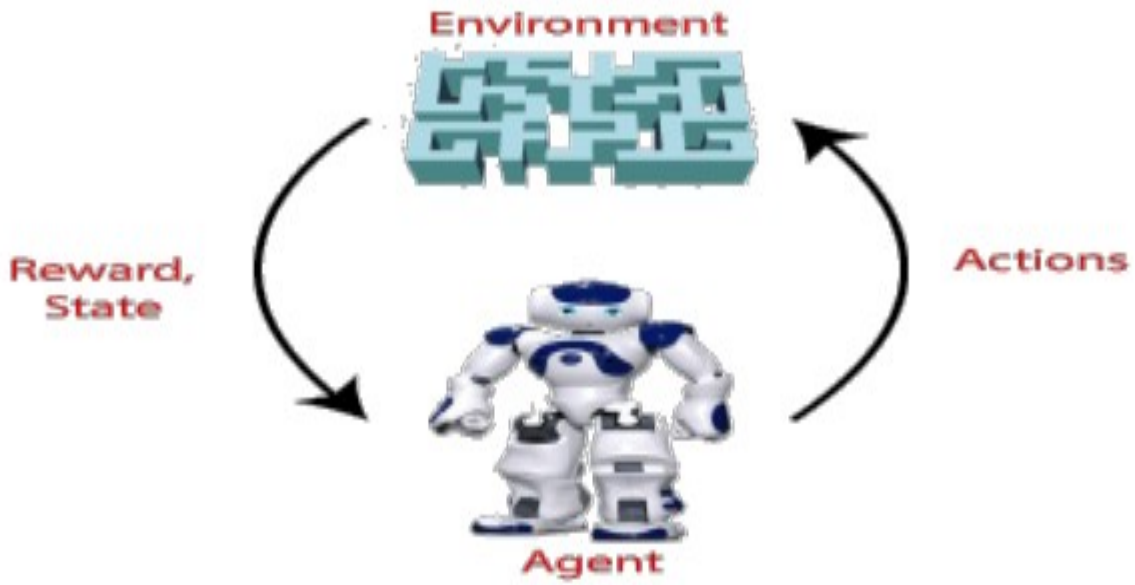
- Unsupervised learning is **intrinsically more difficult** than supervised learning as it does not have corresponding output.
- The result of the unsupervised learning algorithm might be **less accurate as input data** is not labeled

Re-Inforcement Learning

- Reinforcement learning involves an agent learning to make decisions by interacting with an environment.
- In Reinforcement Learning, the agent **learns automatically using** feedbacks without any labeled data unlike supervised learning.
- The agent receives feedback in the form of rewards or penalties based on the actions it takes.
- Reinforcement learning is applicable when an agent needs to learn a sequence of actions to achieve a goal in a dynamic environment, receiving feedback to guide its learning process.

Re-Inforcement Learning

- The agent interacts with the **environment and explores** it by itself.
- **The primary goal** of an agent in reinforcement learning is to **improve the performance** by getting the **maximum positive rewards.**



Re-Inforcement Learning common terms

- **Agent():** An entity that can **perceive/explore** the environment and act upon it.
- **Environment():** A **situation in which an agent is present or surrounded** by.
- **Action():** Actions are the moves taken by an agent within the environment.
- **State():** State is **a situation returned** by the environment after each action taken by the agent.
- **Reward():** A feedback **returned to the agent** from the environment to **evaluate the action of the agent.**
- **Policy():** is a strategy applied by the agent for the next action based on the current state.
- **Value():** It is expected **long-term retuned** with the discount factor and opposite to the short-term reward.

Key Features of Re-Inforcement Learning

- In RL, the agent **is not instructed** about the environment and what actions need to be taken.
- It is based on the **hit and trial** process.
- The agent takes **the next action and changes states** according to the feedback of the previous action.
- The agent may get **a delayed reward.**

Key Features of Re-Inforcement Learning

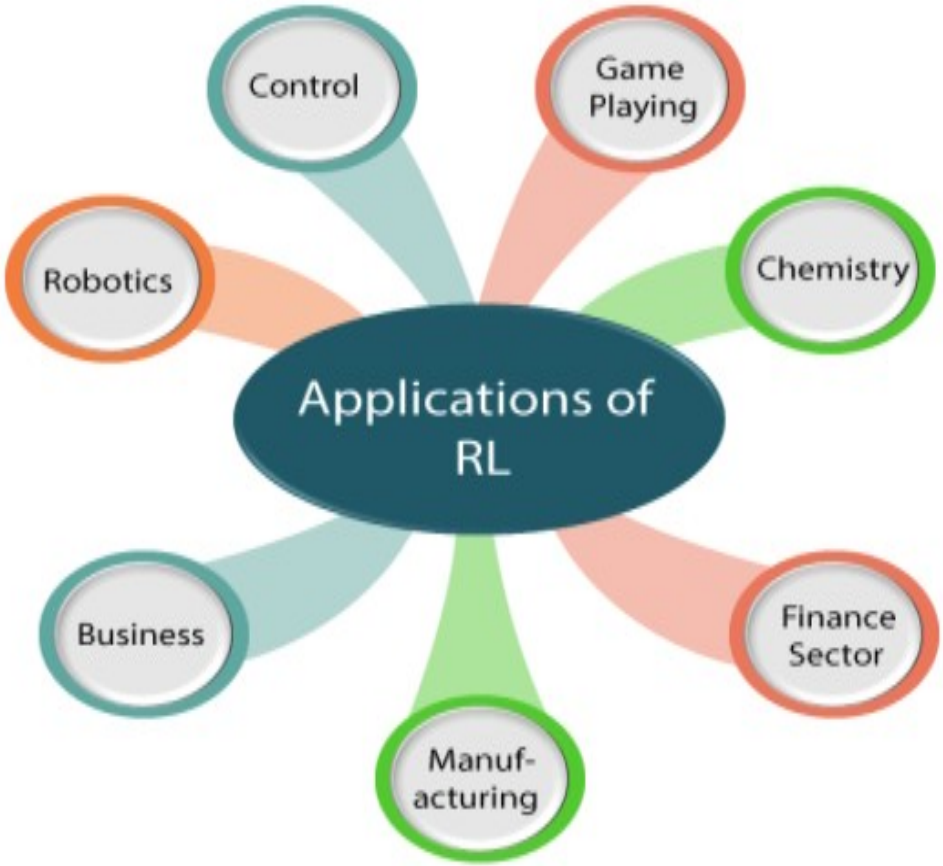
There are mainly two types of reinforcement learning, which are:

- Positive Reinforcement
- Negative Reinforcement

Positive Reinforcement: The positive reinforcement learning means **adding something to increase the tendency that expected behaviour would occur again.**

Negative Reinforcement: The negative reinforcement learning is opposite to the positive reinforcement as it increases the **tendency that the specific behavior will occur** again by **avoiding the negative condition.**

Application areas of RL



Supervised ML- Popular Machine Learning Algorithms

Linear Regression

- one of the most popular and simple machine learning algorithms
- used for predictive analysis ,predictions for continuous numbers such as **salary, age**
- show **linear relationship** between the dependent and independent variables

The equation for the regression line is: $y= a_0+ a*X$
Here, y= dependent variable, X= independent variable, a_0 = Intercept of line.
Linear regression is further divided into two types:

- **Simple Linear Regression:** In simple linear regression, a single independent variable is used to predict the value of the dependent variable.
- **Multiple Linear Regression:** In multiple linear regression, more than one independent variables are used to predict the value of the dependent variable.

Linear Regression example

Senario -1. Assume some company x spent the following cost for advertisement and get sale values as indicated ? **And then , the company wants to do the advertisement of $200 in the year 2019 and wants to know the prediction about the sales for this year.**

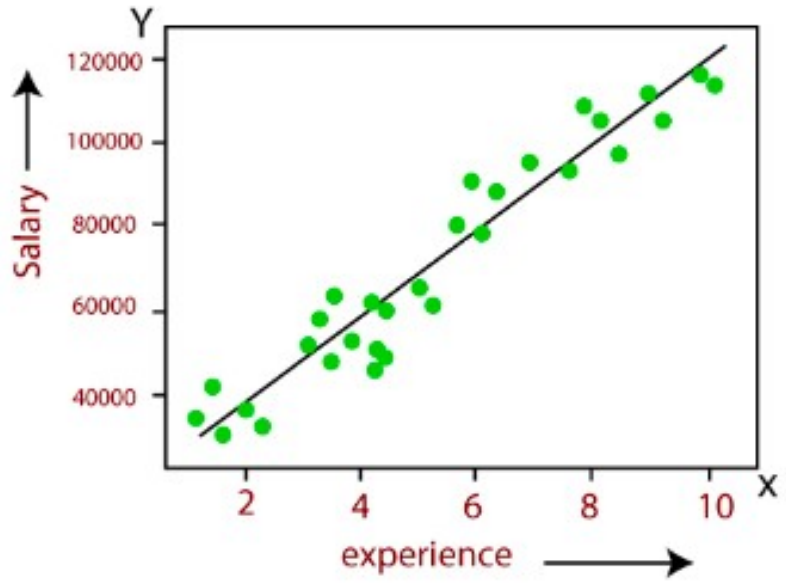
Example :

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Linear reg cont...

Senario 2.company wants prediction of the salary employees based on Experience

Here we are predicting the **salary of an employee** on the basis of the year of experience.



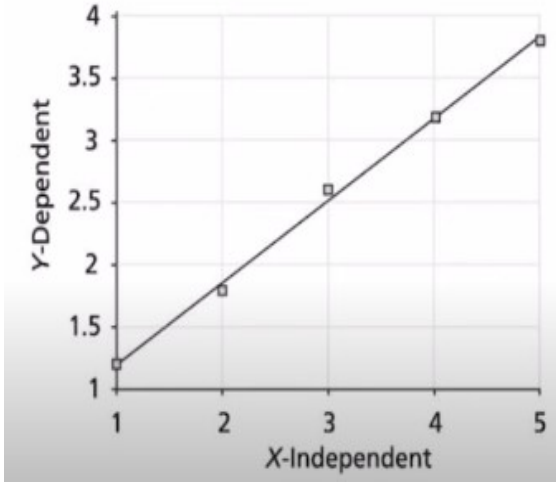
LR problem cont..

Senario 3.company x wants **prediction of the sales** based on different weeks ,wants to know what will be at week 7^{th} or 12^{th} **?**

- Let us consider an example where the five weeks' sales data (in Thousands) is given as shown in Table.
- Apply linear regression technique to predict the 7^{th} and 12^{th} week sales.

x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

Just finding the best fitting line



x_i (Week)	y_j (Sales in Thousands
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

Formula

$$y = \alpha + \beta x$$

β = slope

α = y-intercept

y = y- coordinate

x = x-coordinate

- Linear regression equation is given by
- $y = a_0 + a_1 * x + e$

- *where*
- $a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$
- $a_0 = \bar{y} - a_1 * \bar{x}$

So/n cont...

- Here, there are 5 items, i.e., i = 1, 2, 3, 4, 5.

	x_i (Week)	y_j (Sales in Thousands)	x_i^2	$x_i * y_j$
	1	1.2	1	1.2
	2	1.8	4	3.6
	3	2.6	9	7.8
	4	3.2	16	12.8
	5	3.8	25	19
Sum	15	12.6	55	44.4
Average	$\bar{x} = 3$	$\bar{y} = 2.52$	$\overline{x^2} = 11$	$\overline{xy} = 8.88$

- *where*
- $a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$
- $a_0 = \bar{y} - a_1 * \bar{x}$

Get correct regression line

- $\bar{x} = 3$ $\quad \bar{y} = 2.52$ $\quad \overline{x^2} = 11$ $\quad \overline{xy} = 8.88$

- $a_1 = \frac{(\overline{xy})-(\bar{x})(\bar{y})}{\overline{x^2}-\bar{x}^2} = \frac{8.88-3*2.52}{11-3^2} = 0.66$
- $a_0 = \bar{y} - a_1 * \bar{x} = 2.52 - 0.66 * 3 = 0.54$
- **Regression equation is**
- $y = a_0 + a_1 * x$
- $y = 0.54 + 0.66 * x$

Linear Regression

- **Regression equation is**
- $y = a_0 + a_1 * x$
- $y = 0.54 + 0.66 * x$
- The predicted 7th week sale (when x = 7) is,
- **y = 0.54 + 0.66 x 7 = 5.16**
- the predicted 12th week sale (when x = 12) is,

Application of Linear Regression

Some **popular applications** of linear regression are:

- Analysing trends and sales estimates
- Salary forecasting
- Real estate prediction

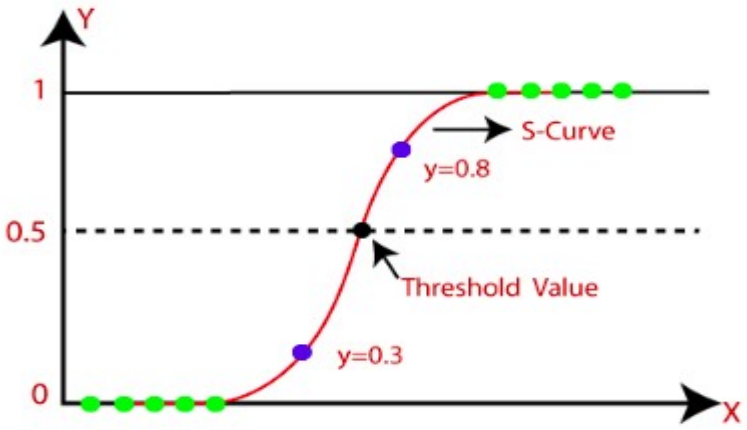
Terminologies Related to the Regression Analysis

- **Dependent Variable**: The main factor in regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable.**
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a predictor.
- **Outliers:** Outlier is an **observation which contains either very low value or very high value** in comparison to other observed values. so it should be avoided.
- **Multicollinearity:** If the **independent variables are highly correlated with each other** than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while **ranking the most affecting variable.**
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting.** And if our algorithm **does not perform well even with training dataset,** then such problem is called **underfitting.**

Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms
- the outcome must be a **categorical or discrete value.** It can be either Yes or No, 0 or 1, true or False, etc.
- it gives **the probabilistic** values which lie **between 0 and 1.**

In Logistic regression, instead of fitting a regression line, we fit **an "S" shaped logistic function,** which predicts two maximum values (0 or 1).



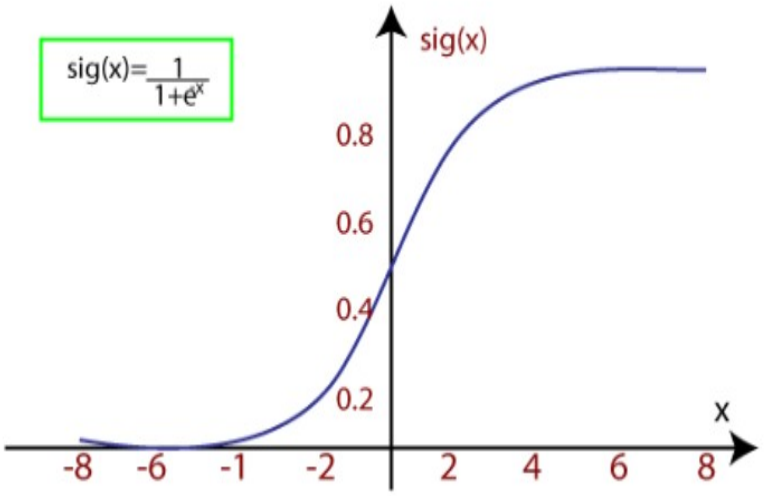
The S-form curve is called the Sigmoid function or the logistic function.

Types of Logistic Regression

- On the **basis of the categories**, Logistic Regression can be classified into three types:
- **Binomial**: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- **Multinomial**: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- **Ordinal**: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Logistic Regression cont..

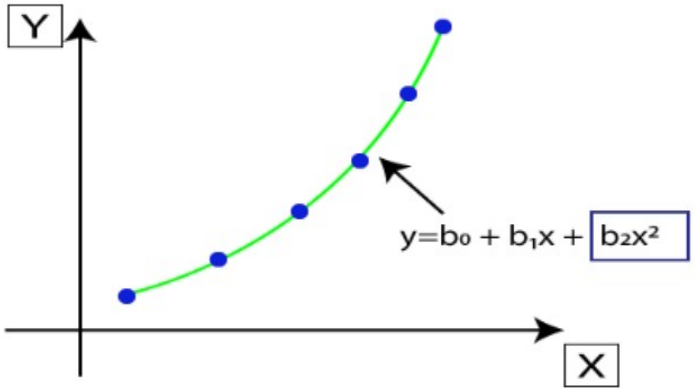
- It is a predictive analysis algorithm which works on the concept of probability.
-



f(x)= **Output between 0 and 1 value,**
x= **input to the function** and
e= **base of natural logarithm.**

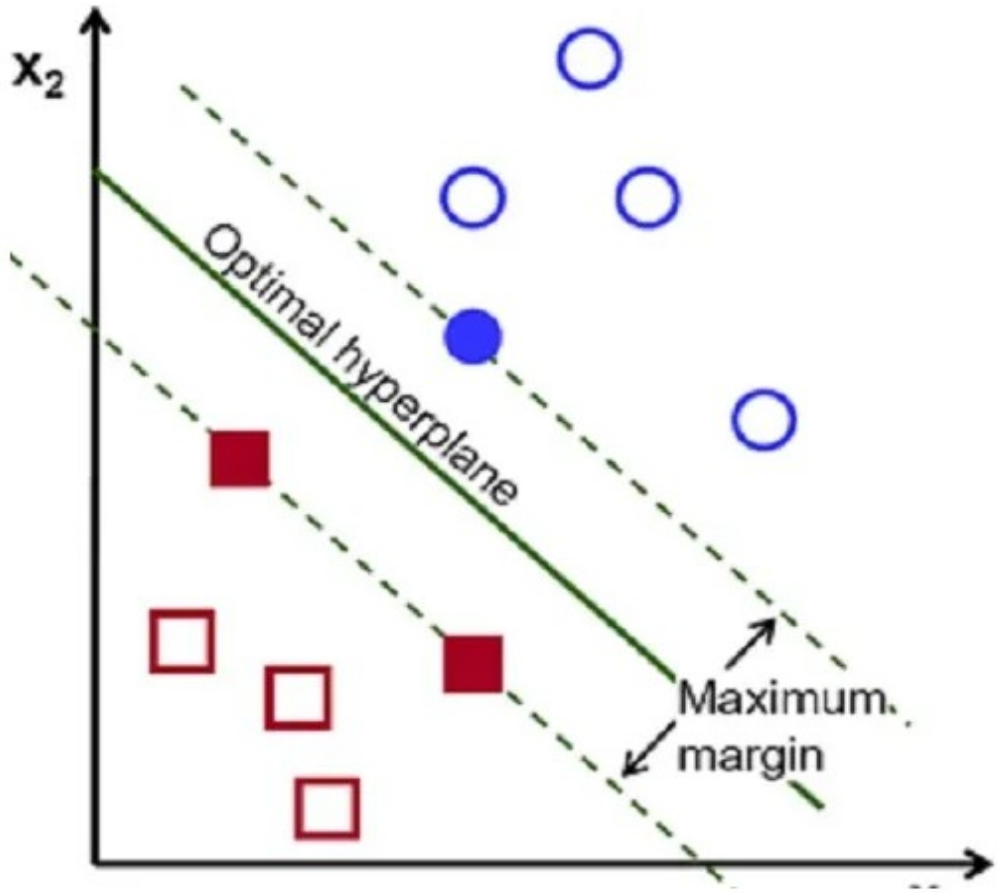
Polynomial Regression:

- Polynomial Regression is a type of regression which **models the non-linear dataset using a linear model.**
- **Tries to capture non-linear r/p b/n independent and dependent variable**
- its a **non-linear curve** between the conditional values of y



➢The equation for polynomial regression also derived from linear regression equation that means Linear regression equation $Y= b_0+ b_1x$, is transformed into Polynomial regression equation $Y= b_0+b_1x+ b_2x^2+ b_3x^3+.....+ b_nx^n$.

SVM Machine Learning algorithm



SVM

- **Support Vector Machine (SVM)** is one of the **most useful supervised ML** algorithms.
- It can be used for both classification and regression tasks.

Basic idea of support vector machines

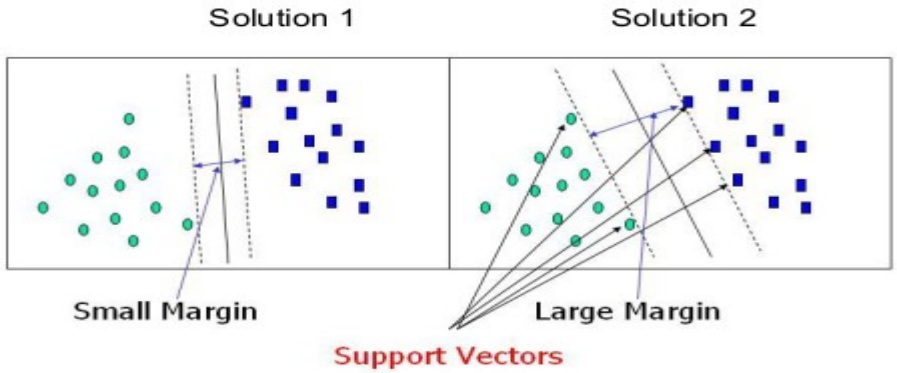
- SVM is a **geometric model** that views the input data as two **sets of vectors** in an n-dimensional space.
- It constructs **a separating hyperplane in that space,** one which **maximizes the margin between the two data sets.**

SVM Machine Learning algorithm

- A good separation is achieved by the **hyperplane** that has the **largest distance** to the neighbouring data points of both classes.

- **The vectors (points)** that **constrain the width of the margin** are the **support vectors**.

- **Support vectors** are the data points that **lie closest to the decision surface**

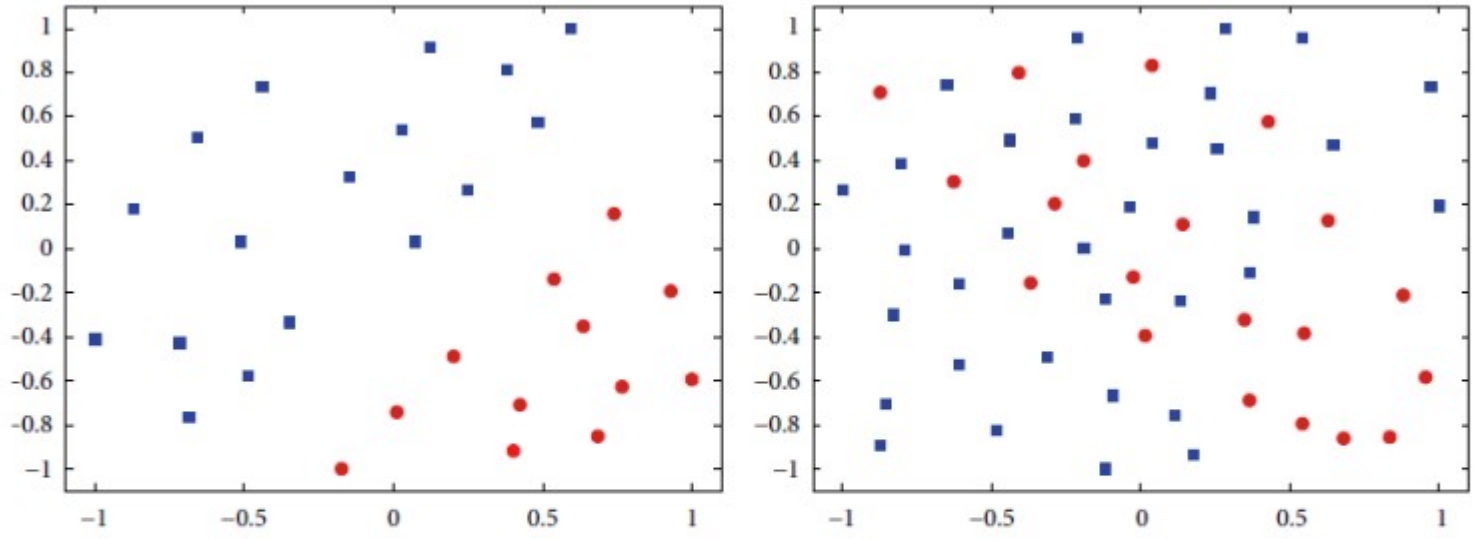


An SVM analysis finds the line (or, in general, hyperplane) that is oriented so that the **margin between the support vectors is maximized.**
In the figure above, Solution 2 is superior to Solution 1 because it **has a larger margin.**

SVM Terminologies

- **Kernel**: It is a function used to map **a lower-dimensional data into higher dimensional data.**
- **Hyperplane**: it is **a separation line between two classes**
- **Boundary line**: Boundary lines are **the two lines apart from hyperplane,** which **creates a margin for data points.**
- **Support vectors:** Support vectors are the datapoints which are nearest to the hyperplane and opposite class.

SVM Machine Learning algorithm

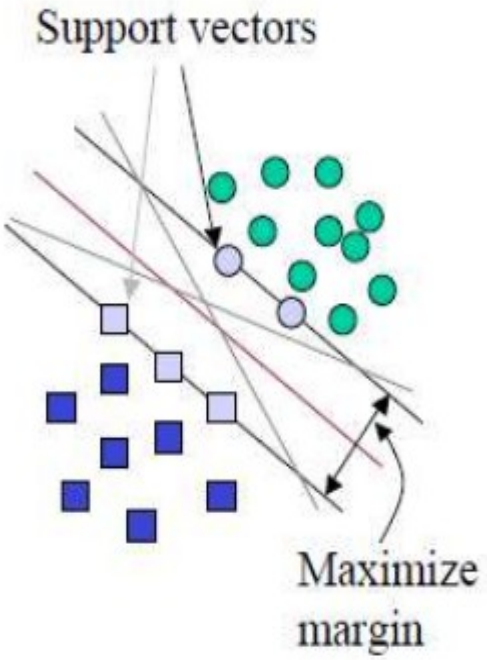


Which one is easy to separate?

SVM Machine Learning algorithm

SVMs maximize the **margin around the separating hyperplane.**

- The decision function is fully specified by **a subset of training samples,** the support vectors.
- 2-Ds, it's a **line.**
- 3-Ds, it's a **plane.**

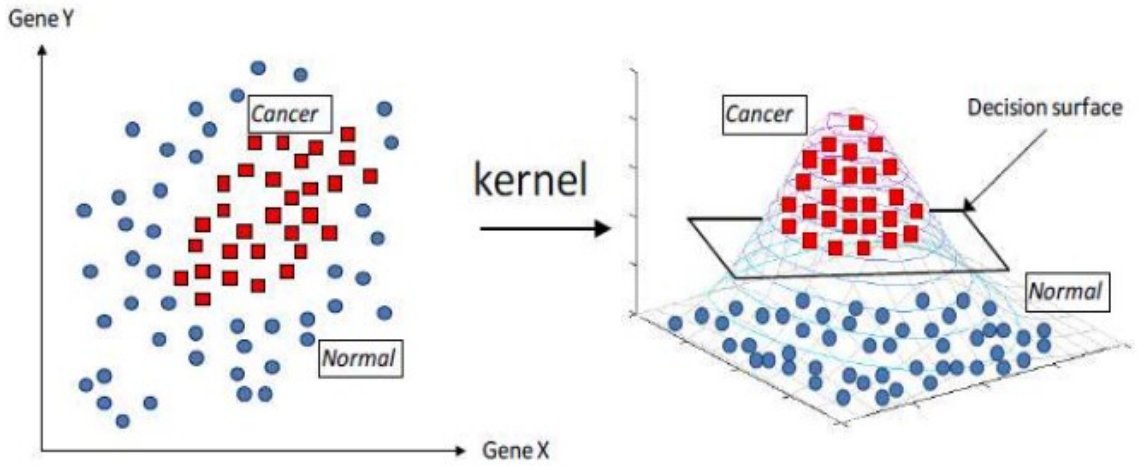


SVM Machine Learning algorithm

– **hyperplane** for linearly separable patterns

-A hyperplane is a **linear decision surface** that splits the space into two parts

– For non-linearly separable data-- t**ransformations of original data to map** into new space – the Kernel function



SVM Machine Learning algorithm

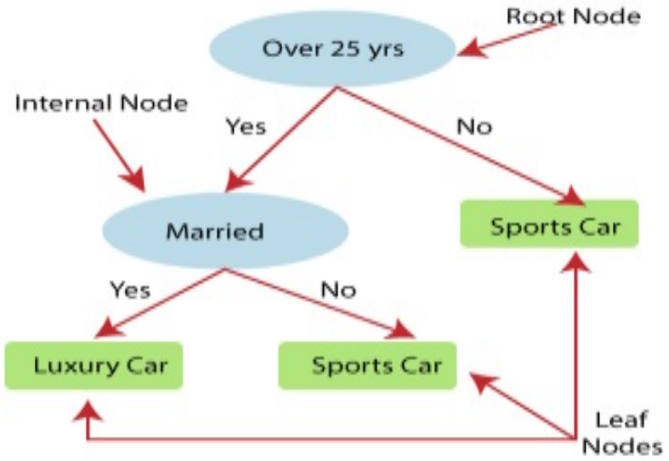
Advantage of SVM:

- Robust to very large number of **variables and small samples**
- Can learn both simple and highly complex classification models
- Employ sophisticated mathematical principles to **avoid overfitting**
- Can be used for **both classification and regression** tasks

-Effective in cases of limited data.

Decision Tree Regression

- **Decision Tree** is a supervised learning algorithm which can be used for solving both **classification and regression** problems.
- It can solve problems for both **categorical and numerical** data
- Ex: the following model is **trying to predict the choice of a person** between Sports cars or Luxury car.



Decision Tree Regression

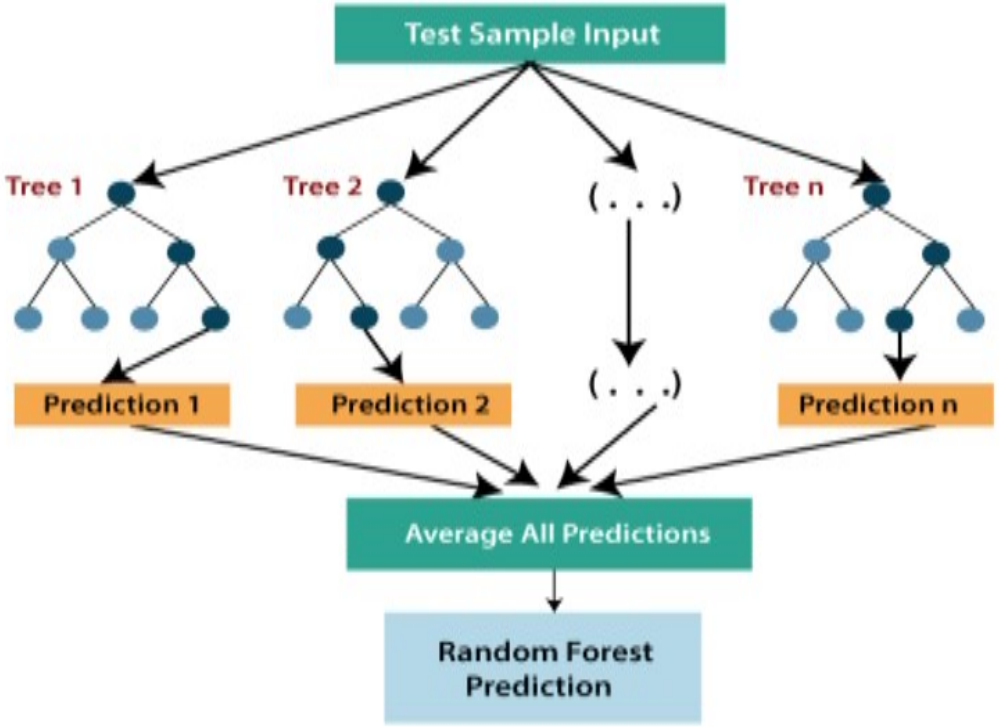
- Decision Tree regression builds **a tree-like structure** in which each **internal node represents the "test"** for an attribute, each **branch represent** the **result of the test,** and each **leaf node** represents the **final decision or result.**
- A decision tree is constructed starting from **the root node/parent** node (dataset), which splits into **left and right child nodes** (subsets of dataset).
- These child nodes are further divided **into their children node,** and themselves become the **parent node of those nodes.**

Random Forest

- Random forest is one of the most powerful supervised learning algorithms which is capable of performing **regression as well as classification** tasks.
- The Random Forest regression is an **ensemble learning method** which **combines multiple decision trees** and predicts the final output based on the **average of each tree output.**
- The **combined decision trees** are called as **base models,** and it can be represented more formally.
- $g(x) = f0(x) + f1(x) + f2(x) + \ldots$

Random Forest

- With the help of **Random Forest regression,** we can **prevent Over-fitting** in the model by **creating random subsets** of the dataset.

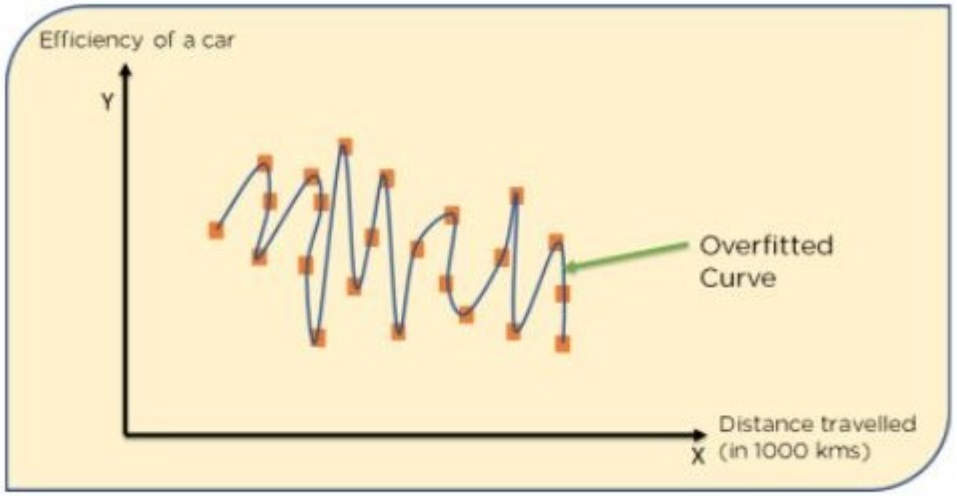


Regularization

- While training a machine learning model, the **model can easily be overfitted or under fitted.**
- if we allow our **machine learning model** to look at **the data too many times,** it will find **a lot of patterns in our data,** including the ones which **are unnecessary and tries to fit each data point on the curve is called Overfitting.**

Regularization

- **Overfitted** model

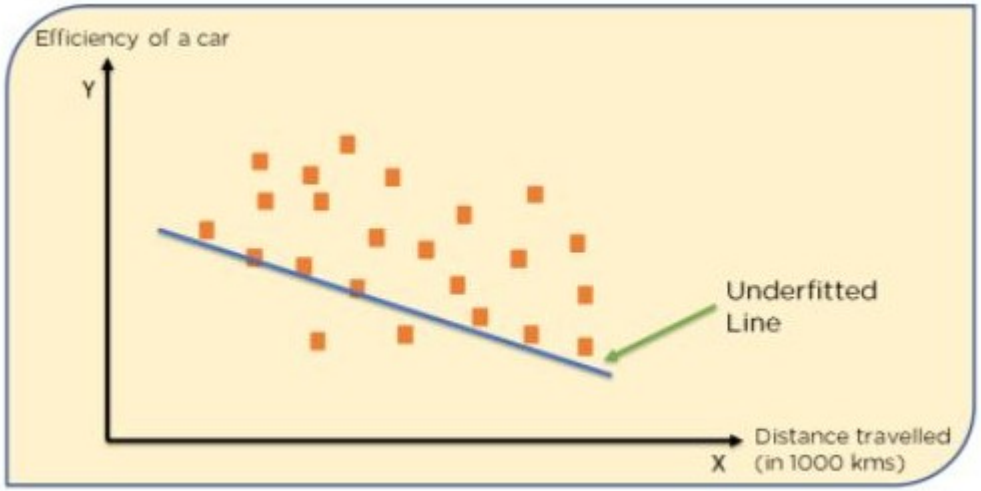


- in a scenario where the model **has not been allowed** to look at our data a **sufficient number of times,** the model **won't be able to find patterns** in our test dataset.

- A scenario where a machine learning model can **neither learn the relationship between variables** in the **testing data nor predict or classify a new data point** is called Under fitting

Regularization

- **under-fitted** model

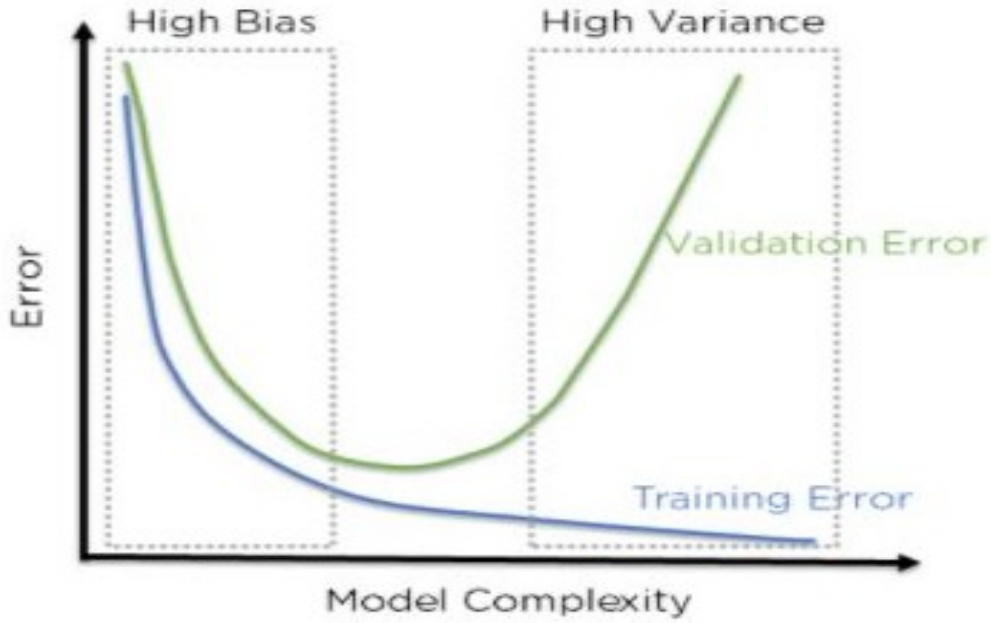


Bias vs Variance

- **A Bias occurs when an algorithm** has **limited flexibility** to learn from data. Such models **pay very little attention to the training data** and oversimplify the model
- therefore the **validation error or prediction error will occur**
- Such models always lead to a high error on **training and test data.**
- **High Bias** causes **underfitting** in our model.
- **Variance** defines the **algorithm's sensitivity to specific sets** of data.
- A model with a high variance pays a **lot of attention to training data**

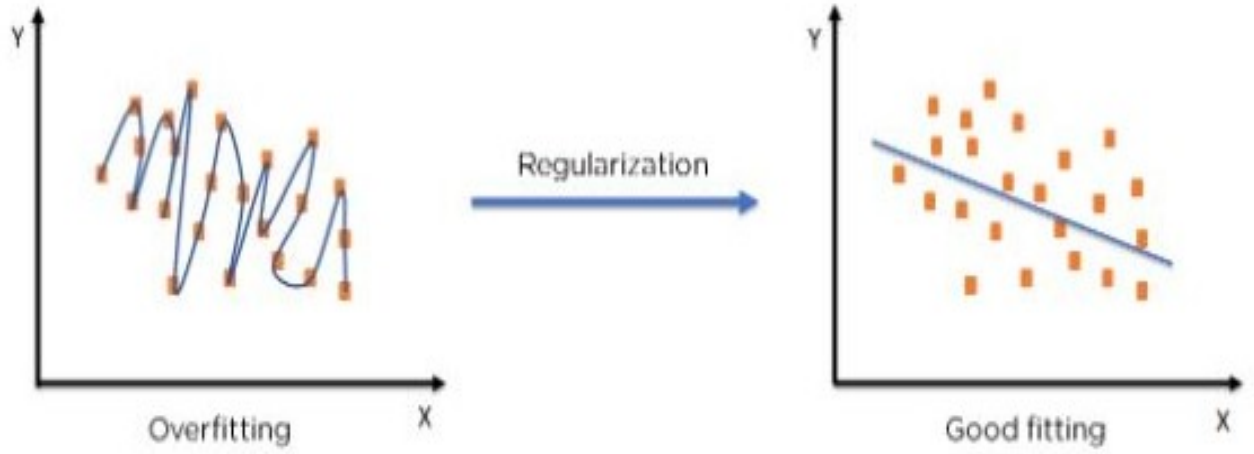
Bias vs Variance

- Variance **does not generalize** therefore the **validation error or prediction error** are **far apart** from each other.
- Such models usually **perform very well on training data** but have **high error rates on test data.**



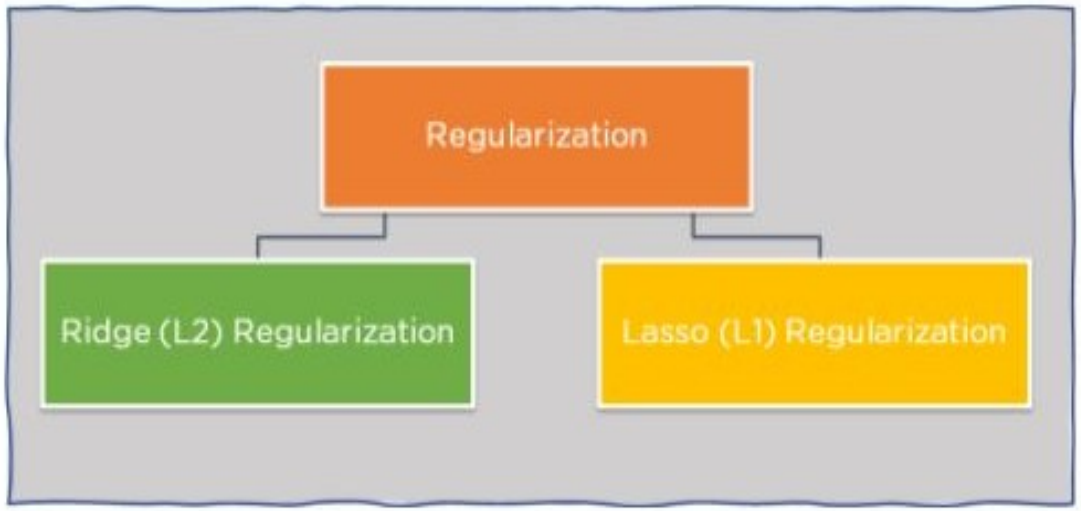
Regularization in Machine Learning

- an **optimal model** is one in which the model is **sensitive to the pattern in our model,** but at the same time can **generalize to new data.**
- Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.
- prevent **overfitting or underfitting**



Regularization Techniques

- There are two main types of regularization techniques: Ridge Regularization and Lasso Regularization.
- Both are regularized version of Linear regression algorithm



Ridge Regression(L2) regularization

- Also known as Ridge Regression, it modifies the **over-fitted or under fitted models** by adding the **penalty equivalent to the sum of the squares** of the magnitude of coefficients.
- Ridge regression is a regularization technique, which is used to r**educe the complexity of the model.** It is also called as **L2 regularization.**
- It helps to solve the problems **if we have more parameters**

Cost function = Loss + $\lambda \times \sum \|w\|^2$

Here,

Loss = Sum of the squared residuals

λ = Penalty for the errors

W = slope of the curve/ line

The equation for ridge regression will be:

$$L(x, y) = \text{Min}\left(\sum_{i=1}^{n} (y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n} (w_i)^2\right)$$

Lasso Regression(L1) regularization

- It modifies the over-fitted or under-fitted models by **adding the penalty equivalent** to the sum of the **absolute values of coefficients.**
- It is similar to the Ridge Regression except that penalty term **contains only the absolute weights instead of a square of weights.**
- **Also known as L1**

Cost function = Loss + λ x $\sum\|w\|$

Here,

Loss = Sum of the squared residuals

λ = Penalty for the errors

w = slope of the curve/ line

$$L(x, y) = \text{Min}\left(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n} |w_i|\right)$$

Cost function

- **Cost function:** Its purpose is to **quantify the difference between** the **predicted values** of a model and the **actual values** observed in the data.
- evaluates how well the **model's predictions match the actual target** values.
- In linear regression, the cost function is often the **Mean Squared Error (MSE) or Mean Absolute Error (MAE),** which measures the average squared or absolute difference between the **predicted and actual values.**
- In logistic regression, the cost function is typically the **Log Loss or Cross-Entropy Loss**

Logistic Regression example

```python
# Importing required libraries
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix

# Loading and preparing the data
data = pd.read_csv('data.csv')
X = data[['feature1', 'feature2', '...']]  # Selecting predictor variables
y = data['outcome']  # Selecting outcome variable

# Splitting the data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Creating and fitting the Logistic Regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Making predictions on test set
y_pred = model.predict(X_test)

# Evaluating the model
accuracy = accuracy_score(y_test, y_pred)
confusion = confusion_matrix(y_test, y_pred)

# Printing the results
print(f'Accuracy: {accuracy}')
print(f'Confusion Matrix: {confusion}')
```

Exercise -work on prediction

There is a car making company that has recently launched a new car,has data as follow

So the company **wanted to check predict whether a user will purchase the product or not, one needs to find out the relationship between Age and Estimated Salary.**

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1

Source of data

https://www.kaggle.com/code/sandragracenelson/logistic-regression-on-user-data-csv/input

Lab Requirements

1. Install Python3
2. Install Anacoda Distribution
3. Use Jupitor Notebook
4. Install some packages using pip package manager using the editor
   - -pandas
   - -Sklearn
   - -matploit lib
   - -numpy