Machine Learning !!

Definition !!

"Study of Computer Algorithms that allows the Computer Program to automatically improve through Experience"

[Training]

[Efficient - learning]

By Tom Mitchel (founder of ML Department)
School of Computer Science at Carnegie Mellon University.

* "ML is teaching the machine about something"

How !!

1. Collect and clean the data
2. Algorithm (model) — select (Readymade) / Built.
3. Teach the model essential pattern from data (Training).
4. Expect the model to give helpful Answer.

Ex To Design a System that determine from MRI Scan, whether Tumor is present or not.

(1) Collect large No of MRI reports

Ex 10000 MRI report < 6000 has Tumor
4000 do not have Tumor } Labelled Data.

Test 30%. Train 70%.

(2) Build an efficient algorithm that detects presence or absence of Tumor in an MRI Scan.
[Expert Consultation – Radiologist].

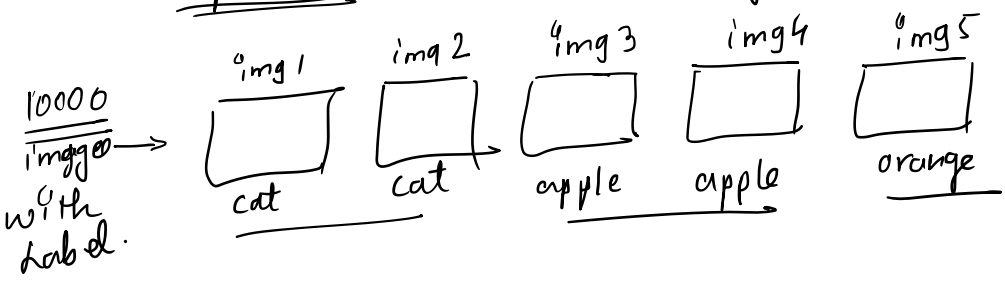
(3) To the Algorithm feed the 6000 MRI Scans and allow the model to learn (train).

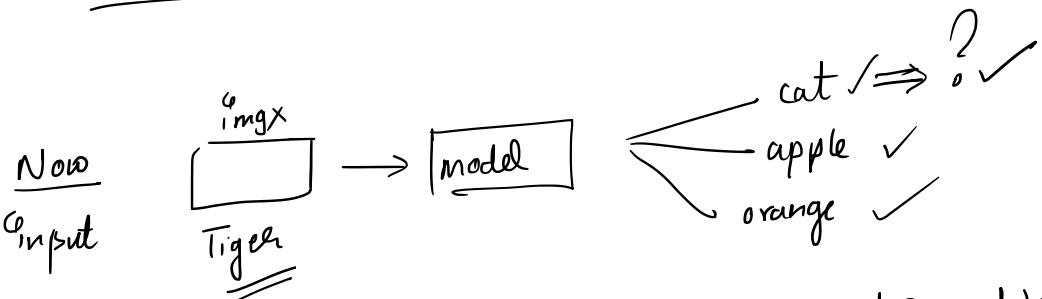
(4) Use around 3000 Images (MRI Scans) for Testing

(5) Use this model to determine presence or absence of Tumor from a New Image (MRI Scan)

Types of machine learning Algorithms.

① Supervised machine learning Algorithms.

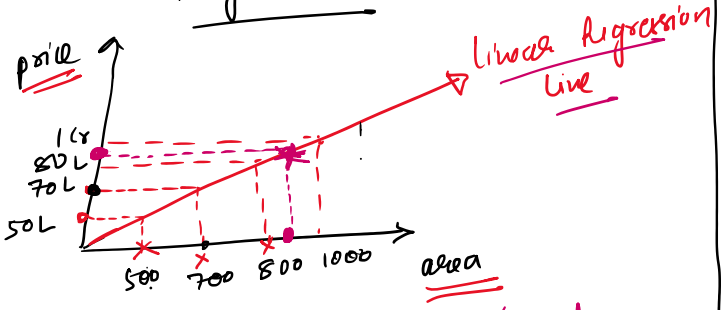




* With lot of images as input, model will be able to identify pattern and will be able to predict.

Supervised.

Regression.



Ex for area = 875 (input)

Price = 8750000 (output)

Here o/p is continous value.

Here area ⇒ Independent variable

price ⇒ Dependent variable

Classification.

Here o/p is not continous value but could be Boolean or some class / category as output.

* Spam detection → Spam Hai / Spam Nahi Hai

* MRI Scan For Tumor Detection → Tumor Hai / Tumor Nahi Hai

In above cases the o/p belongs to

Here area ⇒ Independent variable

price ⇒ Dependent variable

Here the dependent and Independent Variable can have Continous value.

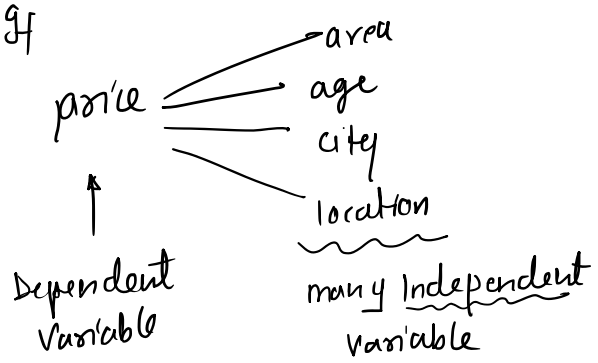
Types ⇒

if price → area.

Single Independent variable

→ Simple Regression

if

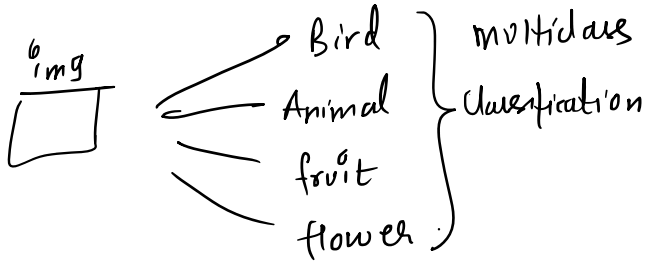


When we have more than one Independent variable ⇒ multiple Regression

In above cases the o/p belongs to one of the two dasses.

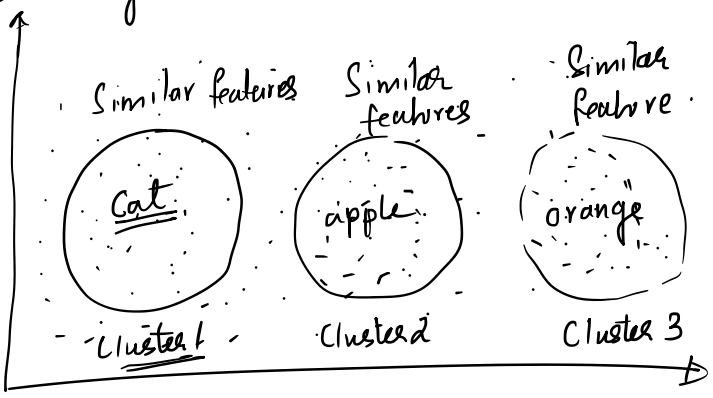
→ Binary classification.

if o/p can be one of many dasses (more than two) then it is called as multiclass classification.



Unsupervised Learning

Given 10000 images and no label are specified.



* Here the input is only data and no labels are associated with data.
* Not Sure about type of output
* Unsupervised Algo will work on 10000 images and will create cluster of images based on the similarities.
* It is "SELF LEARNING"

Unsupervised.

Clustering

(1) Model will identify the pattern in input data and will create cluster.

Ex Market Segmentation.

* Study the Spending / Income of population and create cluster
  (1) High Earning

Association

(1) We find dependencies of one data item on another.

(2) This dependency will help to map the relation and will enhance prediction.

(3) It is like "If - Then"

Ex Used by Amazon known as market Basket Analysis.

population

(1) High Earning
(2) Medium Earning
(3) Low Earning

ex Used by Amazon
Market Basket Analysis.
"If a person purchases cellphone then the person has tendency to purchase screen guard & backcover"

If → Then

(1) Apriori Algorithm → is a breadth first search based approach which calculates the support bet" items.
ex bread influences the buyer to buy milk and eggs so the mapping helps increases the profit.

(2) FP Growth Algorithm →
→ Frequent pattern Algo finds the count of the pattern that has been repeated.

Application of Unsupervised learning

→ Helps us in understanding pattern which can be used to cluster the data based on various features
→ Understanding various defects in dataset which we are not been able to detect initially.
→ It helps in mapping the various items based on the dependencies of each other.

Real world Appln →

(1) Amazon → uses unsupervised learning to learn the customer purchase pattern and recommend the products that are most frequently bought together (Association Rule Mining)

(2) Credit Card Detection → Can we unsupervised learning Algo to learn about various pattern of users and their usage of Credit Card.

→ If the card is used in parts that do not match the behaviour, an alarm is generated which could possibly be marked fraud

* Disadvantage of UnSupervised learning

(1) There is no way of obtaining the method the data is sorted as dataset is unlabelled

(2) They may be less accurate as input data is not known and labelled by humans.

Semi Supervised.

Text Document Classifier

Ex 1000000 articles. and need to classify them into
- News
- Literature
- Research paper
- medical Report

Label is not provided.

Manually labelling the 1000000 articles not possible.

→ we will label (10000) articles [Supervised Learning]

→ my model will be trained on these 10000 articles and will use the pattern identified to classify the remaining 990000 articles [UnSupervised]

* Uses small amount of labelled data
* and large amount of unlabelled data.
* Benefits of Both Labelled and unlabelled data.
* Overcome the challenge of finding large amount of Labelled data.

Ex ① Speech Analysis → we will label audio files which will need lot of human resource and then we will use some technique that will help to improve traditional speech analysis tool.

② Web Content Classification → Organizing the knowledge available on billions of web pages will need advance processing, but the task needs human intervention to classify the content.

Reinforcement Learning [Experential Learning]

→ Here the agent/model learns how to behave in an Environment by performing action and experiencing the result.

Types

Episodic Learning

→ Here we have Start and End state, thus an episode is created.

→ Thus the feather action is based on feedback of result of earlier action.

Ex Fear of dog after being bitten is Episodic Learning.

Continous Task Learning

* There is no terminal state here
* Agent/model that does Automated Stock Trading goes for Continous Learning

* Challenges for Machine Learning →

① Data Collection → Data plays a key role in any the Core.

* 60% of work of a data scientist does is collecting relevant data
* Repositories like kaggle, ML Repository etc are good only for initial learning
* For real case scenarios data needs to be collected through web-scraping or through clients etc.
* Data Collection of relevant data for specific problem is challenge.

② Less Amount of Training Data →

Two important thing that we do while doing a ML project

① Select a learning Algo
② train the model with acquired data

Finding Sufficient amount of relevant data for specific problem is challenge

(3) Poor Quality of Data →

We cannot use the collected data directly for training. As the data might not be ready for training & we need to clean the data.

→ Data Preprocessing needs to be done by filling missing values, extract and rearrange what model needs

(4) Irrelevant / Unwanted Features →

Eg To predict no of hours a person needs to Exercise →

[Collected Need → age, gender, weight, height and location]

→ Remove / filter Unwanted features.

(5) Overfitting the Training Data →

→ Over generalizing is something that we do very frequently. This is known as Overfitting

Avoid →

① Gather more training data.
② Select a model with fewer features.
③ Fix data errors.

(6) Underfitting the Training Data →

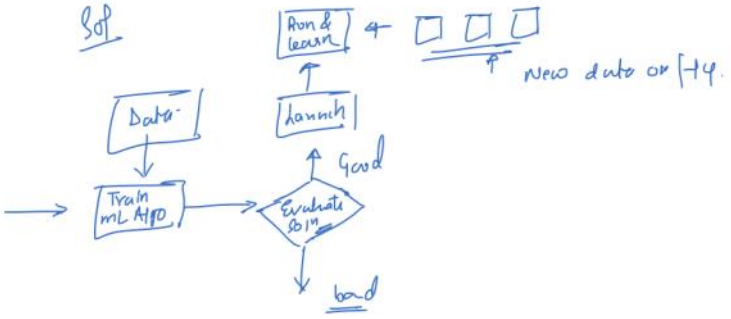
→ It happens when model is too Simple to Understand the base Structure of data.
→ When we have less Information to Construct Model.

Avoid ① Need to feed better features to learning Algo
② Remove noise from data.

(7) Offline Learning & Deployment of Model

→ Challenges in Deployment

1. Lack of practise & dependency issues
2. low understanding of underlying model with business.
3. Understanding of business problems.
4. Data Available online might be irrelevant so we need to create source.



7 stages of ML are :->

(1) Problem Definition -> understand the problem that someone will solve

Questions :
(1) What is the business
(2) Why does problem needs to be solved
(3) What is measurable goal?
(4) If probabilistic in nature, then does available data allow to model it

(2) Data Collection -> Identify the sources

Questions -> (1) What kind of data do I need
(2) Where is the data available
(3) How can I obtain it
(4) How to store and access it efficiently.

(3) Data Preparation → Can take from 70 – 90% of the work time.

Steps Involved like →

- Data filtering
- Data Validation & Cleansing
- Data Formatting
- Data Aggregation & Reconciliation.

(4) Data Visualization → perform Exploratory Data Analysis (EDA)
→ helps in identifying pattern.

Ways of Visualization →

- Area chart • Bar chart • Bubble cloud
- Heat Map. Histogram.

W/o 1. Diagram, word cloud etc

(5) ML Model → Here magic Happens

→ Finding of pattern in data using supervised or Unsupervised.
→ Clustering, Regression, classification etc.
→ Apply mathematical model to train ML algorithms that will help to make predictions.

(6) Feature Engineering → It is collections of methods for identifying an optimal set of inputs to ML Algorithms.

Characteristics of Good features

(1) Data represented must be unambiguous.
(2) Ability to capture linear & non linear relationship among data points
(3) Capture context detail.

(7) Model Deployment → to put the model in production environment to make data driven decisions in automated way.

ML models can be deployed as

① SaaS (Software as Service) ≡

→ cloud Application servid
→ Uses Internet to deliver appln which are managed by third party vendor to its user
→ Executes on browser and do not requires downloads or installation on client side

Ex → Google Apps, Dropbox, PackMy cloud Salesforce

② PaaS (Platform as a Service)

→ provides cloud components to certain s/w used for appln.
→ It delivers a framework for developers that can built upon and use to create Customized Applications.

Ex → AWS Elastic, RedHat Openshift, Windows Azure. Vmware, IBM Blue mix

③ IaaS (Infrastructure as a Service)

→ Cloud Infrastructure services known as IaaS are made of highly scalable and Automated computing Resources.
→ It is fully self Service for accessing & Monitoring things like Compute, Storage, N/w and other infrastructure related Service on pred-are on demand basic.

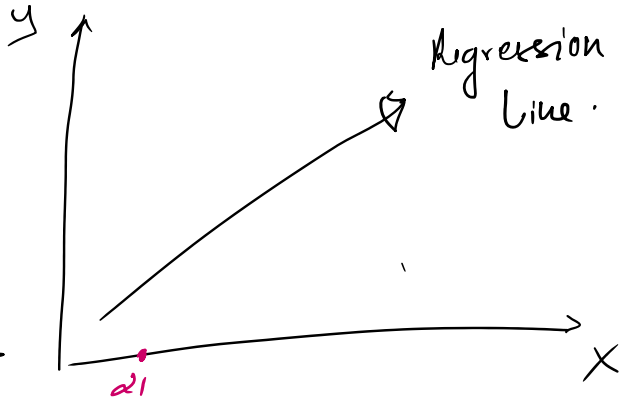
Ex → AWS, Azure, GCP (Google Cloud Platform), IBM Cloud.

	INDEPENDENT	DEPENDENT	x^2	xy
SR.NO	X	Y		
1	43	99	1849	4257
2	21	65	441	1365
3	25	79	625	1975
4	42	75	1764	3150
5	57	87	3249	4959
6	59	81	3481	4779
	$\Sigma x = 247$	$\Sigma y = 486$	11409	20485

↓ Actual value. $n = 6$

Now predict if $X = 55$, what is y ?

Σx^2 Σxy



→

$$y = a + bx + e$$

- y : Dependent variable
- a : Intercept
- b : decides impact of X on y.
- e : Error

[if x↑ y↑ +ve Impact]
[if x↑ y↓ -ve impact].

a = intercept
b = slope (coefficient of Independent Variable)
c = Error

Let us assume $e = 0$

$$y = a + bx \quad \text{— (1)}$$

To find a and b.

Take Σ on both side of 1

$$\Sigma y = \Sigma a + \Sigma bx$$

$$\Sigma y = a\Sigma 1 + b\Sigma x$$

$$\Sigma y = an + b\Sigma x \quad \text{— (2)}$$

multiply Eq 2 with Independent Variable X.

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{— (3)}$$

multiply Eq 2 with independent ...

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad -(3)$$

Here $\frac{Eq\ 2}{Eq\ 3}$

$$486 = a6 + b247 \quad \checkmark$$

$$20485 = a247 + b11409 \quad \checkmark$$

$$a = 65.14$$

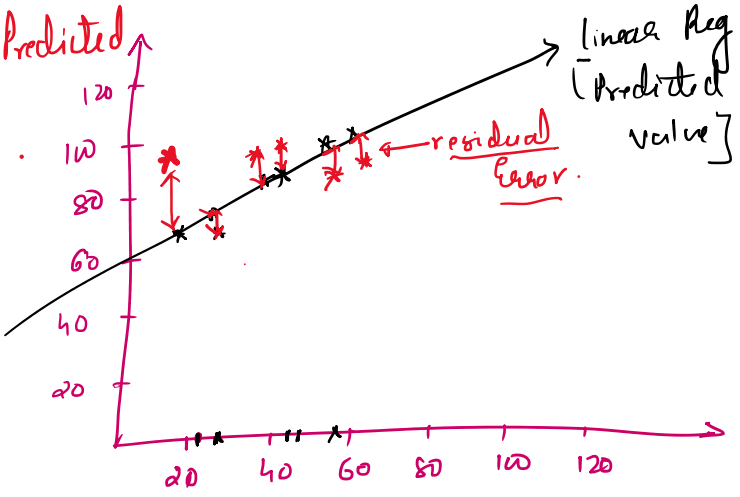
$$b = 0.385$$

$$y = 65.14 + 0.385x$$

Using this let us calculate Predicted Y for Every X

Actual − Predicted

SR.NO	INDEPENDENT X	DEPENDENT(ACTUAL) Y	PREDICTED Y	DIFFERENCE
1	43	99	81.695	-17.305
2	21	65	73.225	8.225
3	25	79	74.765	-4.235
4	42	75	81.31	6.31
5	57	87	87.085	0.085
6	59	81	87.855	6.855



Residual Error = Error between the Actual value and Predicted value.

Yaad Rakho

$$y = a + bx$$

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

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\ldots}$$

Quantitative Analysis

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

Given X & Y pair $\quad \boxed{y = a + bx}$

Find a, b.

If more than one independent Variable ->

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 \cdots$$

b_1: impact of x_1 on y

b_2: impact of x_2 on y

b_3: impact of x_3 on y.