PAGE No.
DATE / /

Machine Learning (ML) is a subset of AI.

Q. What is ML?

⇒ ML is about extracting knowledge from data.

* It is also known predictive analytics or statistical learning.
* It involves many fields such as statistics, computer science, mathematics and artificial intelligence.
* It has many applications into our day-to-day life, such as, movie recommedations in several apps, what food to order or which product to buy.
* Also, it has applications in research oriented field, finding distant planets, understanding galaxy, discovering new particles, analyzing DNA sequences and providing personalized cancer treatment.
* More examples, filtering spam emails, voice-to-text messages.

* ML is usually divided into two categories.

Predictive / supervised	descriptive / Unsupervised learning / knowledge discovery
Goal: To learn mapping from input x to output y	Goal: Find interesting patterns in the data. [knowledge discovery]
$D = \{(x_i, y_i)_{i=1}^{N}\}$ is a training set.	* Much less well-defined problem. so there is no obvious error metric to use.
N is no. of training example.	
Each x_i — vector of numbers	

representing features, attributes or covariates. The values are stored in matrix form called as design matrix.

The output y_i [response variable] is a categorial or nominal variable from some finite set $y_i \in \{1, 2, \ldots C\}$. This known as classification / pattern recognition. When y_i is real valued the problem is known as regression.

* There are two more types of ML.
  (3) Semi-supervised ML.
  (4) Reinforcement ML.

* Def^n of ML: Arthur samuel (1959):
  Field of study that gives computers the ability to learn without being explicitly programmed.

PAGE No.
DATE / /

1. Supervised ML.

Classification — Output variable y_i is categorical.

Regression — Output variable y_i is continuous.

algorithms
- Random forest alg.
- Decision Tree alg.
- Log

* ML is all about determining patterns - analyzing training data in such manner that the trained algorithm can perform tasks that the developer didn't originally programmed it to do.

* What is big data ?

⇒ Big data is substantially different from being just a large database. Big data implies lot of data, but it also includes the idea of complexity and depth.

Eg. self-driving cars.

Inputs : Few HD cameras & couple hundred sensors that provide information at a rate of 100 times/s. ⇒ Raw input data that exceeds 100 Mbps. [Processing that much data is incredibly hard].

* The acquisition of big data is also daunting.

Imp: how the dataset is stored & transferred so that the system can process it.

PAGE No.
DATE / /

In most cases, developers try to store the dataset in memory to allow fast processing Using a hard drive to store the data would be too costly, time-wise.

* bias: The bias is characteristic of simpler algorithms that can't express complex mathematical formulations.

Limits of bias / or bias present in the problem
1. In collecting data - eg. of U.S. poll in 1936
2. In method
3. In algorithm.

*

* Hypothesis

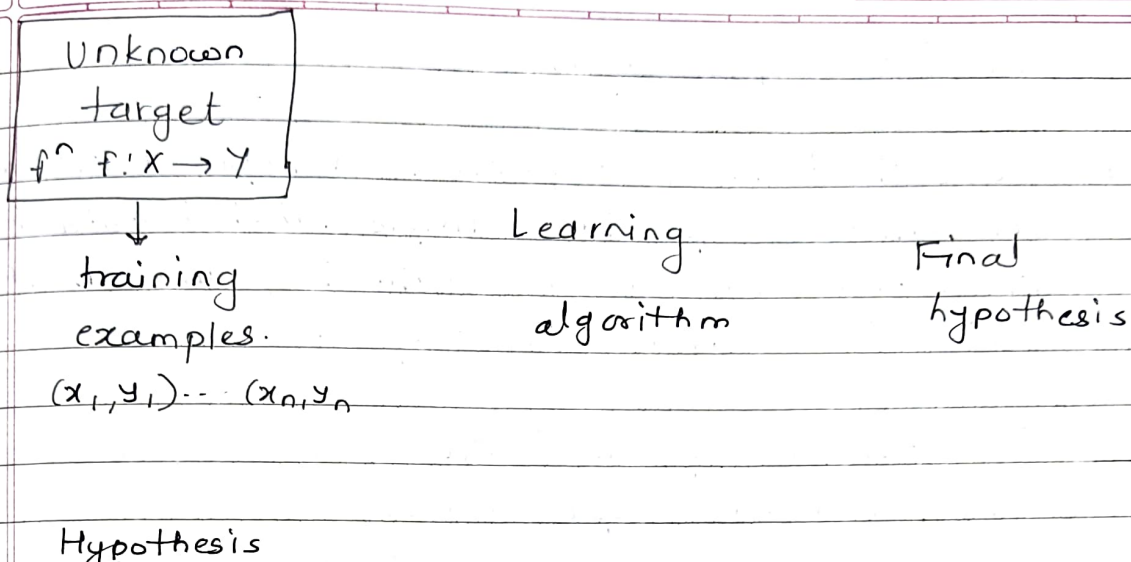
ML professionals conduct experiments that aim to solve a problem & they make an initial assumption for the solution of the problem. This assumption in ML is known as hypothesis.

The hypothesis is defined as the supposition or proposed explaination based on insufficient evidence or assumptions. It is just a guess based on some known facts which is yet to prove..
A good hypothesis is testable, which results in either true or false.

Eg: Based on tropical meterogy data scientist claim that tommorow may be a rainfall.
hypothesis.

* Hypothesis space (H): Hypothesis space is defined as a set of all possible legal hypotheses. It is also known as a hypothesis set.

PAGE No.
DATE / /



* In supervised ML, techniques, the main aim is to determine the possible hypothesis out of hypothesis space that best maps input to the corresponding / correct outputs.

** Hypothesis: It is defined as the approximate function that best describes the target in supervis ML algorithms.
It is mainly based on data as well as bias & restrictions applied to data.

$y = mx + c$ ← one of the hypothesis.

* Inductive bias: Inductive bias is the ability of the machine learning algorithm to generalize beyond the observed training examples to handle unseen

PAGE No.
DATE / /

Def^n of types of ML.

① Supervised ML : Sup. ML is a type of ML method in which we provide sample labeled data to the ML system in order to train it, and on that basis, it predicts the output.

The goal of supervised ML is to map input data with the output data.

Further classified into two categories of alg.

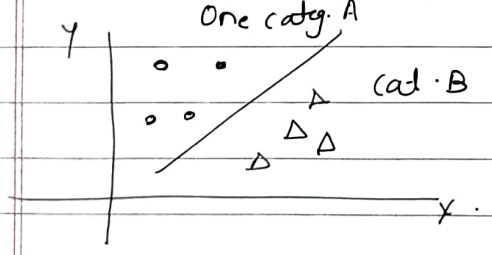
① Classification ② Regression.

↓

The classification algorithm is a supervised ML technique that is used to identify the category of new observation on the basis of training data.

In classification, a program learns from the given dataset / observations and then classifies new observation into a number of classes or groups.
eg. Yes or No, Spam or No spam. etc.

* The output variable is a category
i.e. $y = f(x)$ y - categorial output



Types of Classifications :

① Binary classifiers : If the classification problem has only two possible outcomes, then it is called Binary classifier. Eg. Yes / No.

PAGE No.
DATE / /

(2) Multi-class classifier: If a classification problem has more that two outcomes, then it is called as multi-class classifier.
eg. Classification of types of music / movies.

Types of ML classification algorithms.

(1) Linear models.
(1) Logistic regression
(2) Support Vector Machines.

(2) Non-linear models.
(1) K-Nearest neighbours.
(2) Kernel SVM.
(3) Naive Bayes
(4) Decision Tree Classification.
(5) Random Forest classification.

(2) Regression: It is a supervised ML technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.
* It is mainly used for prediction, forecasting, time series modeling, ie market trend.
* Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints & the regression line is minimum.

PAGE No.
DATE / /

Terminologies related to regression ML.

(1) Dependant variable: The main factor in Regression analysis which we want to predict is called the dependant variable (target variable).

(2) Independent variable: The factor which are used to predict the values of the dependent variables are independent variable (predictor).

(3) Outlier: It is an observation which contains either very low value or very high value in comparison to other observed values. It may hamper the result, so it should be avoided.

(4) 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 variables.

(5) 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 out algorithm does not perform well even with the training dataset then such problem is called underfitting.

* To avoid overfitting:
① Cross validation ③ Removing features
② Training with more data ④ Early stopping
⑤ Regularization ⑥ Ensembling.

PAGE No.
DATE / /

Types of regression sup. ML:

(1) Linear regression.
(2) ~~Logistic regression~~
(3) Polynomial -11-
(4) Support vector regression.
(5) Decision tree -11-
(6) Random forest -11-
(7) Ridge regression.
(8) Lasso regression.

① Linear regression:

(1) Used for predictive analysis.

* Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).
* If there is only one input variable (x), then such linear regression is called simple linear regression. And if there are more than one input variable, then such linear regression is called multiple linear regression.

$Y = mX + c$.

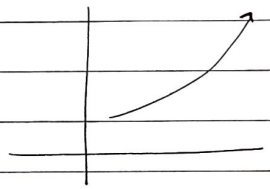
m & c are called linear coefficients.

eg. Salary forecasting.
Real estate prediction.
Analyzing trends and sales estimates.

PAGE No.
DATE / /

(2) Polynomial regression:

* It models the non-linear dataset using a linear model.
* It fits a non-linear curve between the value of x and y.
* In this regression, the original features are transformed into polynomial features of given degree & then modeled using a linear model.



$y = b_0 + b_1 x + b_2 x^2$.

regression coefficients.

(3) Support vector regression:

Bias

* In the realm of ML, there are many biases like selection bias, overgeneralization bias, sampling bias, etc.

Inductive bias is important factor; without which learning will not be possible.

There are some assumptions made about the data itself in order to make learning possible is called as inductive bias.

Def of Inductive bias: The inductive bias of a learning alg is a set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered.

PAGE No.
DATE / /

* ~~Linear regression, Data cleaning~~

* Induction bias

(1) Induction ~~bias~~ is the process of learning something general from something specific. [opposite is called deduction].

(2) A Bias is a preference for one course of action over another.

(4) Induction would be impossible without such a bias, because observations may generally be extended in a variety of ways.

eg. of inductive bias:

(1) The linear model presupposes that each of the input characteristic & the target have a linear connection.

The no free lunch theorem of ML demonstrates that there is no single best bias & that, whenever algorithm A outperforms algorithm B in one set of problems, there is an equal no. of problems in which alg. B outperforms A.

Or, finding the optimum model for a ML problem doesn't include searching for a single "master alg" but rather seeking for an alg. with biases that make sense for the issue at hand.

Thm "If all true functions are equally likely then no learning algorithm is better than other".

(5) Predictions for new ~~scean~~ scenarios could not be formed if all of these options were treated equally.

PAGE No.
DATE / /

Explaination of inductive bias is by maths terminology

Consider the problem of learning mapping $f \in F = Y^X$ from an input space X to an output space Y, given a set of training data $D = \{(x_1, y_1), \cdots (x_n, y_n)\} \subseteq X \times Y$. A learning algorithm A takes D as an input and produces a function f (model, hypothesis) f as an output : $f = A(D)$.
The set of candidate models are those functions $f: X \to Y$ that are consistent with the data i.e. $f(x_i) = y_i$ for $i = 1, 2, \cdots n$.
The inductive bias can be understood as the collection of all properties of the learning algorithm responsible for choosing the model $f = A(D)$ as a generalization instead of any other consistent function.

Note: (1) For practical learning algorithm, this defⁿ has limited usefulness. As it is often not possible to give a precise characterization of the inductive bias.
(2) ML algorithms normally operate in a stochastic environment where observations can be noisy, erroneous and inconsistent. Under these conditions, it is not reasonable to insist on consistency.

* A learning algorithm does not normally consider all mappings $f \in F$ as candidate models, but only subset $H \subseteq F$ of these mappings.

PAGE No.
DATE / /

* Cross-validation in ML.

It is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data.

Steps of cross-validations are:

(1) Reserve a subset of the dataset as a validation set.
(2) Provide the training to the model using the training dataset.
(3) Next, evaluate model performance using the validation set. If the model performs well with the validation set, perform the further step, else check for the issues.

* Limitations of cross-validation:

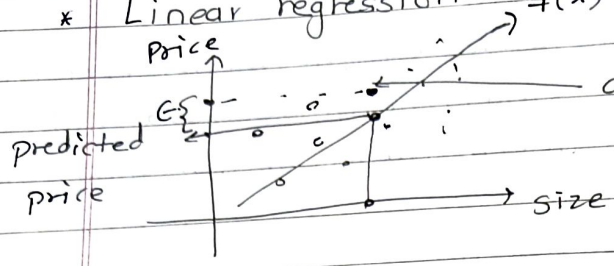
(1) For ideal conditions, it provides the optimum output. But for the inconsistent data, it may produce a drastic result.
(2) In predictive modeling, the data evolves over a period due to which, it may face the differences between the training set and validation sets.

* Applications:

(1) This technique can be used to compare the performance of different predictive modeling methods.
(2) In medical-research field.

PAGE No.
DATE / /

* Linear regression — $f(x)$ — f^n - relationship bet^n house size & price.



actual price. $y \approx f(x)$.

$y_i = f(x_i) + \epsilon_i$ ϵ_i is given as error in prediction for i^{th} house.

ϵ_i is positive, negative or zero for different eg.s

Expected value for error is zero.

$E(\epsilon_i) = 0$.

i.e. y_i can be above or below the line equally likely.

In regression — the tasks

(i) Select the type of function for the given problem. is it linear, qudratic, cubic or any higher order.

(ii) Find the estimated function for the given problem.

* We have to fit a line..

$f(x) = B_1 x + B_0$.

The prediction value or output is

$y_i = B_1 x_i + B_0 + \epsilon_i$

B_1 & B_0 are coefficient of regression model

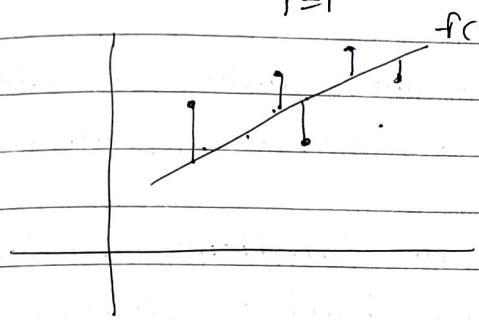
B_1 ↑ slope, B_0 ↑ intercept.

* Residual sum of square RSS is the sum of squared difference of actual value & predicted value.

$$RSS = (y_1 - (B_1 x_1 + B_0))^2 + (y_2 - (B_1 x_2 + B_0))^2 + \cdots$$
$$(y_n - (B_1 x_n + B_0))^2.$$

PAGE No.
DATE / /

$$= \sum_{i=1}^{n} (y_i - (\beta_1 x_i - \beta_0))^2.$$

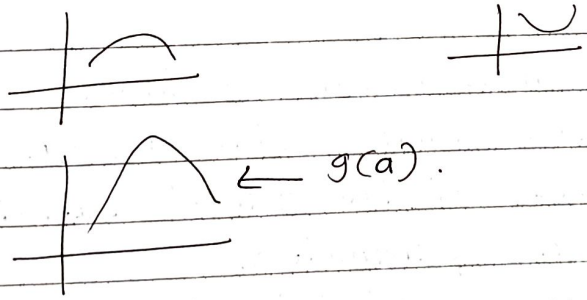


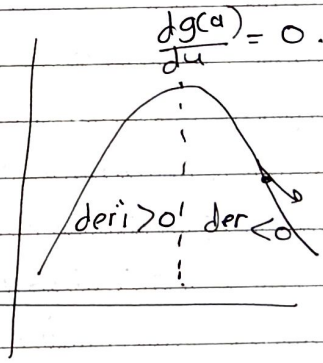
Goal : for regression is to minimize the RSS.

* Try to find the estimated parameters $\hat{\beta}_1$ & $\hat{\beta}_0$.
These are the best guess to actual parameters.

$$\min_{\beta_1, \beta_0} \sum_{i=1}^{n} (y_i - (\beta_1 x_i + \beta_0))^2.$$

* Concave & Convex f^n.





PAGE No.
DATE / /

Gradient descent algorithm:

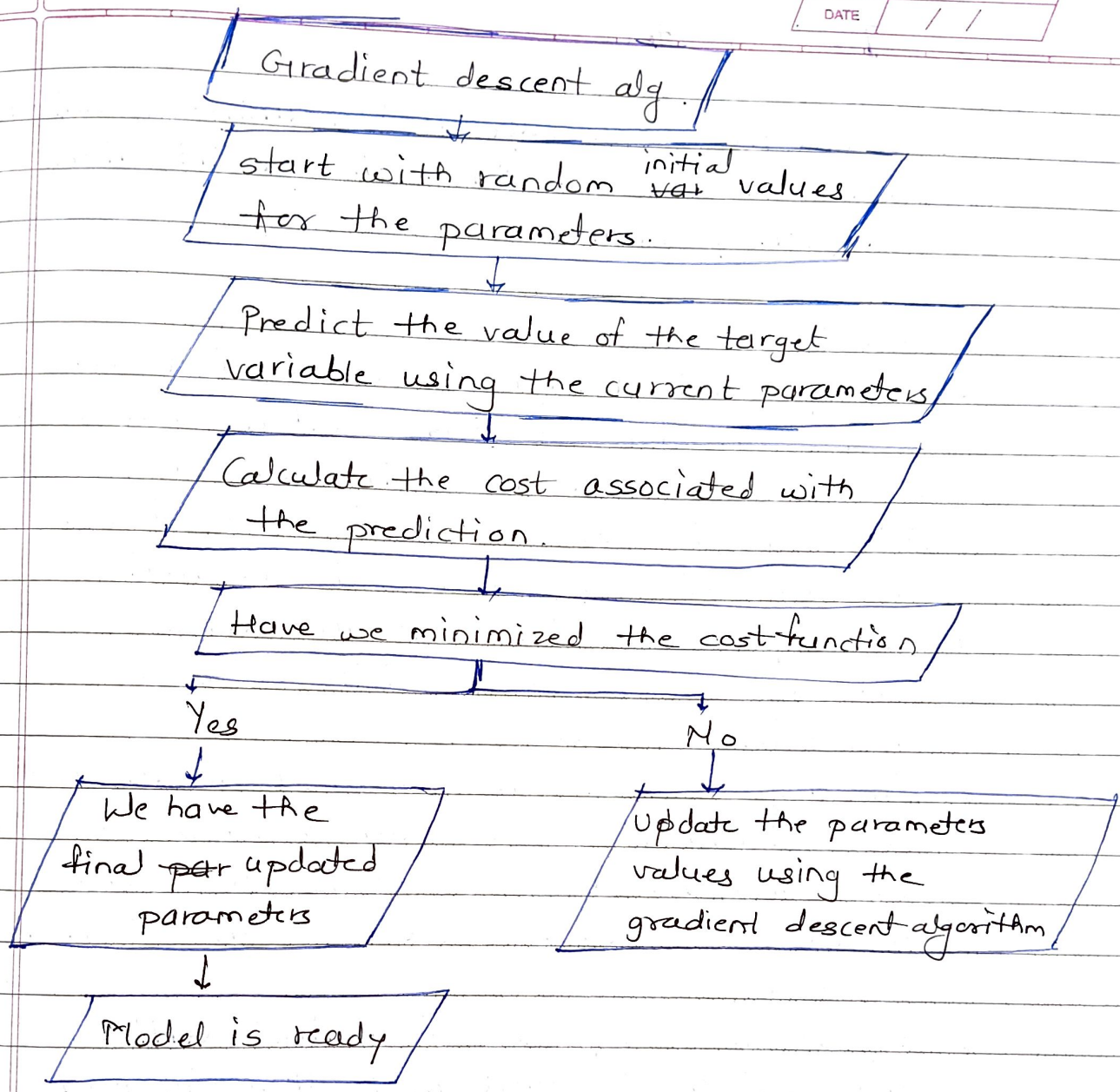
* Gradient descent is a machine learning algorithm that operates iteratively to find the optimal values for its parameters.
* The algorithms considers the function's gradient, the learning rate and the initial parameter values while updating the parameter values.

Q. Why do we need the Gradient descent algorithm?
⇒ The goal of the Gradient descent algorithm is to find the best parameter values that reduces the cost associated with predictions. In order to do this, initially start with random values of these paramters & try to find the optimal ones. To find the optimal values, the gradient descent alg. is used.

Q. How does the gradient descent algorithm work?
⇒ 1. Start with random initial values for the paramters.
2. Predict the value of the target variable using the current parameters.
3. Calculate the cost associated with prediction.
4. Has the minimized cost acheived?

Yes	No
⇒ go to step 6	⇒ go to step 5.

5. Update the parameter values using the gradient algorithm & return the step 2.
6. We have our final parameter.
7. The model is ready.



Q. Obtain the formula of the gradient descent alg. for only one variable ?

Q. -11- two variable. $f(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2$.

PAGE No.
DATE / /

Decision tree

* Decision tree is a supervised learning technique that can be used for both classification & regression problems.

Note:

* A graph G is a triplet consisting of a vertex set V(G), an edge set E(G), and a relation that associates two vertices with an edge.
* If E(G) is ~~an~~ the set of unordered pair of elements of V(G), then G is an undirected graph.
* A graph G is finite if the vertex set V(G), and edge set of G are finite, otherwise G is an infinite graph.
* The degreee of a vertex u in G, is the no. of edges incident to u. [and denoted by $d(u)$].
* The minimum degree of G is the smallest value of the set $\{d(u) : \text{where } u \in V(G)\}$ & it is denoted by $\delta(u)$.
* A path is a sequence of distinct vertices $u_1, u_2, \ldots u_k$ such that u_i is adjacent to u_{i+1} for each i.
  $P_k = \langle u_1, u_2, \ldots u_k \rangle$.
* A cycle is obtained by adding an edge bet^n the end-vertices of a path.
* A cycle on k-vertices has k-edges & it is usually denoted by C_k.
* A graph G is connected, if for any two distinct vertices u & v, there is a path in G from u to v.
* A tree is an undirected graph in which any two vertices are connected by exactly one path.

Or A connected acyclic undirected graph is called a tree.

* Leaf / pendant node: A vertex of a tree which has degree 1 is called leaf node / pendent vertex.
* Internal nodes: All the vertices of a tree with degree more than one are called internal vertices/nodes.
* A forest is an acyclic undirected graph / disjoint union of trees.
* A binary tree is a rooted tree in which every internal node has at most two leaves.

→ Decision trees.

Types: based on the type of target variables.

① Categorical variable decision tree:
Decision tree which has a categorical target variable.

② Continuous variable decision tree:
The decision tree which has a continuous target variable.

* Root node: It represents the entire population or sample & this further gets divided into two or more sets.
* Splitting: It is a process of dividing a node into two or more sub-nodes.
* Decision node: When a node splits into further nodes, then it is called the decision node.
* Pruning: It is the process of removing unwanted branches from the tree.

* Each node in the tree acts as a test case for some attribute, & each edge descending from the node corresponds to the possible answers to the test case.

PAGE No.
DATE / /

x Algorithm for decision tree:

1. Begin the tree with the root node, say S, which contains the complete dataset.
2. Find the best attribute in the dataset using Attribute selection measure (ASM).
3. Divide S into subsets that contain possible values for the best attributes.
4. Generate the decision tree node, which contains the best attribute.
5. Recursively make new decision tree using the subsets of the dataset created in step 3. Continue this process until a stage is reached where you can not further classify the nodes & called the final node as a leaf node.

* The main issue @ comes while implementing a decision tree is that how to select the best attribute for the root node & decision nodes.

⇒ ASM - Attribute selection measure is ~~the~~ used to solve this problem.

Information Gain Gini Index

1. Information Gain:

* It is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.

* [entropy: measure of the randomness of a system]
degree

PAGE No.
DATE / /

* It calculates how much information a feature provides us about a class.
* According to the value of information gain, we split the node & build the decision tree.
* A decision tree algorithm always tries to maximize the value of information gain, & a node/attribute having the highest information gain is split first

Formula

$$\text{Information Gain} = \text{Entropy}(S) - [(\text{Weighted avg}) \times \text{Entropy}(\text{each feature})]$$

$$\text{Entropy}(S) = -P(\text{yes}) \cdot \log 2\ P(\text{yes}) - P(\text{no}) \log 2\ P(\text{no})$$

⇒ The main criteria based on which decision trees $plit.

Imp:
1. Gini impurity: Measures the impurity in a node.
2. Entropy: Measures the randomness of the system.
3. Variance: This is usually used in the Regression model, which is a measure of the variation of each data point from the mean.

* Advantages of decision trees:

1. Easy to explain to others. It does not need any complex mathematical knowledge to understand the result.
2. Can handle irrelevant attributes.
3. Can capture non-linear relationships in the data.
4. Decision tree learning is fast process. It uses greedy algorithms to create trees. Also, prediction by using tree is fast.

PAGE No.
DATE / /

5. Normalization & other cleanning on data is not needed
6. Can handle both numerical and categorical attributes.

* Disadvantage of Decision tree:

1. Decision tree needs to stop somewhere. Otherwise they will give no error on training data & high error on test data.

PAGE No.
DATE / /

* Instance based ML:

1. It is also known as Memory based learning.
2. It is a supervised classification learning alg. that performs operation after comparing the current instances with previously trained instances, which have been stored in memory.
3. It is used in both classification & regression.

* Advantages:

1. Instead of estimating for the entire instance set, local approximations can be made to the target f^.
2. This algorithm can adapt to new data easily.

* Disadvantages:

1. Classification costs are high.
2. Large amount of memory required to ~~to~~ store the data, & each query involves starting the identification of a local model from scratch

* Instance-based algorithms:

(1) K-nearest neighbour.
(2) Self-organizing map.
(3) Learning vector quantization.
(4) Locally weighted learning.
(5) Case-based Reasoning.

PAGE No.
DATE / /

* KNN alg :
- It is a supervised learning alg. & it is used for both classification & regression problems.
- It is also referred to as lazy learning alg. as it does not do any work until it known what exactly needs to be predicted & from what type of variables.
- It is a very efficient cross-validation tool.

Disadvantages associated with memory-based learning are the curse of dimensionality, run time cost scales with training set size, large training sets do not fit into memory, & predicted values in case of regression are not continuous.

* Basis of KNN classification.
1. During the training method, it saves the training examples.
2. During prediction time, k training examples are being found among examples $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$ that are closed to the test example x.
Then, the most frequent class among those y_i's are predicted.

PAGE No.
DATE / /

* Feature reduction:

* It is also known as dimensionality reduction.
* It is the process of reducing the no of features ~~in a resource~~ without losing important information
* Feature reduction can be divided into two processes.
  1. Feature selection
  2. Feature extraction.

Advantages.

* The purpose of using feature reduction is to reduce the no. of features that the computer must process to perform its function. Feature reduction leads to the need for fewer resources to complete ~~co~~ tasks. Less computation time & less storage capacity needed means the computer can do more.

→ ① Feature selection: In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It involves following ways:

1. Filter.
2. Wrapper.
3. Embedded

② Feature extraction: This reduces the data in a high dimensional space to a lower dimensional space.

* Methods of dimensionality reduction:

The various methods for dimensionality reduction include:

① Principal Component analysis (PCA).
② Linear Discriminant analysis (LDA).
③ Generalized Discriminant Analysis (GDA)

PAGE NO.
DATE / /

Advantages of Dimensionality reduction:

(1) It helps in data compression, & hence reduced storage space.
(2) It reduces the computation time.
(3) It also helps to remove redundant features, if any.

Disadvantages.

(1) It may lead to some amount of data loss.
(2) PCA tends to find linear correlations between variables, which is sometimes undesirable.
(3) PCA fails in cases where mean & covariance are not enough to define datasets.

* Note:

Why division with standard deviation?

⇒ Division with the std deviation of the original features ensures that the std deviation of standardized data becomes 1. This will make the importance of all the attributes equal, & our algorithm will not be biased towards any specific feature. [As the goal of PCA is to transform the data such that it retains most of information/variance.

PCA is one such tech for feature reduction.

What is Principle Component analysis PCA?

It is an unsupervised ML technique to reduce the dimensionality of data consisting of a large no. of inter-related features/variable but at the same time retaining as much as possible of the variation present in the original data.

Every attribute/feature in the data is considered as a separate dimension.

PAGE No.
DATE / /

The PCA algorithm transforms the data attributes into a newer set of attributes called Principal components.

For eg. if the data has 10 attributes, then PCA will form a new set of 10 attributes (called Principal components) using the earlier 10 attributes. These ~~new set of~~ principal components are ~~not~~ uncorrelated. & they are arranged such that the first few PCs from the start retain most of the variations present in the original set of attributes.

* Steps in PCA.

(1) Standardization of data:

Calculate the deviation of every element from the mean of the corresponding attributes, ~~i.e.~~ & then divide the result by std deviation of that attribute.

$$Z = \frac{X - \text{mean}(X)}{\sigma(X)}$$

Python $Z1 = (x1 - np.mean(x_1)) / n.p.std(x1)$

$Z_2 =$

$\vdots$

$Z_n =$

(2) Calculating the covariance matrix of data:

Dataset having n attributes will have a $n \times n$ covariance matrix.

$$= \begin{bmatrix} Cov(X,X) & Cov(Y,X) & \cdots \\ Cov(X,Y) & Cov(Y,Y) & \cdots \\ \vdots & \vdots & \vdots \end{bmatrix}$$

PAGE No.
DATE / /

" covariance = np.cov(np.array([z_1, z_2]))

Covariance (X,X) = variance (X,X).
& covariance (X,Y) = covariance (Y,X).

It is a symmetric matrix.
Sign of covariance of (X,Y) tells how the two attributes are related.
If the sign is positive: X & Y will increase or decrease together i.e. they are correlated.
If the sign is negative: If X increases then Y decreases & vice-versa i.e. X & Y are inversely correlated.

3. Calculate Eigenvalue & Eigenvector.

Eigenvectors [unit vectors from the origin] corresponds to the directions along which the PC's will lie.
Eigenvalues are the coeff. attached with eigenvectors that say about the variance carried along with corresponding eigenvectors.

4. Sort eigenvalues in decreasing order
The first PC will be the eigenvector corresponding to the highest eigenvalue. & the second PC will be second eigenvector corresponding to second highest eigenvalue & so on.

Note: For the reduction of dimension, we have can decide to ignore some PC.

What is Scree Plot ?

⇒ The plot of eigenvalues w.r.t. PCs is known as scree plot. This plot is used in determining the no. of PC's to retain. The PC's having eigenvalue > threshold should be retained & the rest can be discarded. This threshold can be considered a hyperparameter & tuned according to the project's needs.

5. Forming the newer feature set along the principal componet axis.
After discarding the unwanted ~~feature~~ components, form a matrix with column entries as eigenvecter.

*

* Where do we use PCA ?
It is used among ML engineers & data scientists. Following are the explicit uses.
(i) Reducing the image size.
(ii) Facial recognition.
(iii) Medical science. [Detecting the correlation betn cholesterol & lipoprotein.

Q. 1) What are principal components in PCA ?
2) What is the need of for PCA in ML projects ?
3) What does a covariance matrix represents.
4) Is PCA affected by the presence of outliers ?

PAGE No.
DATE / /

* Collaborative filtering based recommendations.

What is a recommendation system?

=) A recommendation system generates a compiled list of items in which a user might be interested, in the reciprocity of their current selection of items. It expands user's ~~recommendat~~ suggestions without any disturbance.

For eg, the Netflix recommendation system offers recommendations by matching & searching similar user's habits & suggesting movies that share characteristics with films that users have rated highly / watched.

Types of recommendation systems:

(1) Content filtering recommender systems.
(2) collaborative filtering based recommender systems

(1) Content filtering:

Content filtering expects the side information such as the properties of a song. (Song name, singer name, movie name, language & others).

(2) Collaborative filtering:

It is used by recommendation systems to find similar patterns or information of the users. This technique can filter out items that users like on the basis of the ratings or reactions by similar users.

Types of collaborative filtering.

(1) User-user-based coll. filt.

(2) Item-Item- -ll-

eg.

Users
U1
U2
U3
U4

Rec
can

(1) It

Th
ba
fi
be
+
++

(2) U
users

PAGE No.
DATE / /

eg.

Users	Movie 1	M 2	M 3	M 4
U 1	5	4		5
U 2	4		3	
U 3		1		2
U 4	1	2		

Recommendation systems based on collaborative filtering can be categorized in the following ways:

(1) Item based: This type of recommendation system helps in finding similarities between the items / products. This is done by generating data of the no. of users who bought two or more items together & if the system finds a high correlation then it assumes similarity bet^n products. For eg, there are two products X & Y that are highly correlated when a user buys X, the system recommends buying Y also.

(2) User based: This type of system helps in finding similar users based on their nature of items selection. For eg., one user uses helmet, knee guard & elbow guard, & the second uses only a helmet & elbow guard at then user-based recommendation system will recommend the second user to use a knee guard.

Cosine similarity:

$$sim(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|}$$

PAGE No.
DATE / /

eg item-to-item based collaborative filtering.

User / Item	Item_1	Item_2	Item_3
User_1	2	—	3
User_2	5	2	—
User_3	3	3	1
User_4	—	2	2

* Finding similarities of all the item pairs.

Sim (Item_1, Item_2)
Only user_2 & 3 have rated both items

$I_1 = 5U_2 + 3U_3$.
$I_2 = 2U_2 + 3U_3$ $(5,3) \cdot (2,3)$

$$sim(I_1, I_2) = \frac{5 \times 2 + 3 \times 3}{\sqrt{5^2+3^2} \times \sqrt{2^2+3^2}} = 0.90$$

Sim (Item 2, Item 3)

$I_2 = 3U_3 + 2U_4$
$I_3 = U_3 + 2U_4$

$$sim(I_2, I_3) = \frac{3 \times 1 + 2 \cdot 3}{\sqrt{3^2+2^2}\sqrt{1^2+2^2}} = 0.869$$

$sim(I_1, I_3) = 0.789$

* Generating missing values:

rating of Item2 by User 1.

$$r(U_1, I_2) = \frac{r(U_1, I_1) * s(I_1, I_2) + r(U_1, I_3) \times s(I_3, I_2)}{S(I_1, I_3) + S(I_2, I_3)} = 3.43$$

PAGE NO
DATE / /

$$r(U_4, I_1) = \frac{r(U_4, I_2) \times S(I_2, I_1) + r(U_4, I_3) \times S(I_1, I_3)}{S(I_2, I_1) + S(I_1, I_3)}$$

$$= \frac{2 \times 0.9 + 2 \times 0.789}{0.9 + 0.789} = 2.0$$

Q.1. Given the following data, use PCA to reduce the dimension from 2 to 1.

Feature	1			
x	4	8	13	7
y	11	4	5	14.

=) Step Mean.

$x_i - \bar{x}$ $4 - \bar{x}$ $8 - \bar{x}$

$y_i - \bar{y}$

covariance matrix $= \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$

$\lambda_1 = 30.3849$, $\lambda_2 = 6.6151$

eigen vector $(14 - \lambda_1)u_1 - 11u_2 = 0$

$-11u_1 + (23 - \lambda_1)u_2 = 0$

$\frac{u_1}{11} = \frac{u_2}{14 - \lambda_1} = t$.

$u_1 = t \times 11$, $u_2 = t(14 - \lambda_1)$.

$u_1 = 11$ $u_2 = 14 - \lambda_1 =$

1st week of April

PAGE No
DATE / /

* Logistic regression

* Advantages:

1. Logistic regression performs better when the data is linearly ~~separable~~ independent.
2. It does not require too many computational resources as its highly interpretable.
3. It does not require tuning.
4. It is easy to implement & train ~~the~~ model
5.

	Linear regression	Logistic regression
(i)	Used to solve reg. problem	Used to solve classification problems.
(ii)	The response (dependent) variables are continuous	The response (dependent) variable is categorical
(iii)	It helps to estimate the dependent variable when there is a change in the independent variable.	It helps to calculate the possibility of a particular even taking place.
(iv)	It is a straight line.	It is an S-curve. (S-sigmoid)

* Applications

(1) To predict the weather conditions of a certain place. [sunny, windy, rainy etc).

(2) Ecommerce companies can identify buyers if they are likely to purchase a certain product.

(3) To classify objects based on their features & attributes

1st week of April

Logistic regression, Support vector Machine, Kernel fn & Kernel SVM.

PAGE No.
DATE / /

* Assumptions in a Logistic regression algorithm.

1. In a binary Logistic regression, the dependent variable must be binary.
2. Only meaningful variables should be included.
3. The independent variables should be independent of each other i.e. the model should have little or no multicollinearity.
4. Logistic regression requires quite large sample sizes.

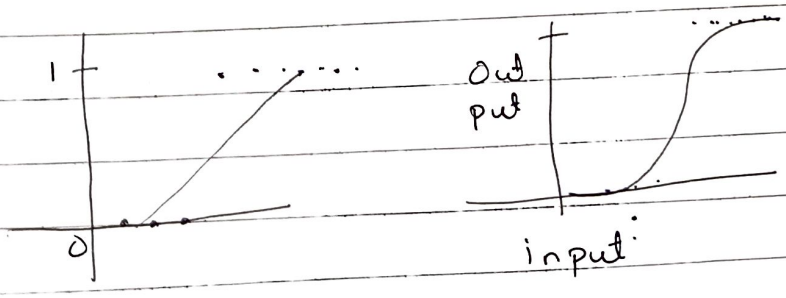
* Types of Logistic regression.

(1) Binomial: In this regression, there are only two possible types of dependent variables such as 0 or 1.

(2) Multinomial: In this regression, there can be 3 or more possible unordered types of the dependent variables.

(3) Ordinal: In this regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium" or "High".

* How does the Logistic regression algorithm work?



2nd Week April — Neural Network, Perception, Multilayer Network, backpropagation, introduction to deep Neural Network, RNN & LSTM

Lets go over the odds of success

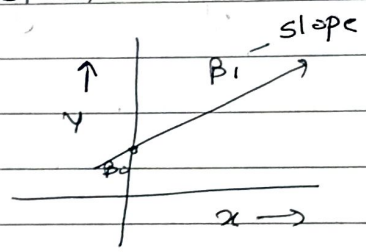
$$\text{odds}(\Theta) = \frac{\text{Probability of an event happening}}{\text{Probability of an event not happening}}$$

$$\Theta = \frac{P}{1-P}$$

Θ ranges from 0 to ∞ & the values of P lies bet^n 0 & 1.

From linear regression.

$$y = \beta_0 + \beta_1 x$$



~~y =~~

$$\log y = \log\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x.$$

$$\frac{P(x)}{1-P(x)} = e^{\beta_0+\beta_1 x}.$$

Componendo

$$\frac{P(x)}{} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

The function $g(z) = \frac{1}{1+e^{-z}}$ is called sigmoid f^n.

It squeezes any real no. to the (0, 1) in open interval. It is less sensitive to outliers unlike ~~to~~ linear regression.

3rd clustering, K-means, adaptive heirarchial clustering, Gaussian mixture model.

PAGE No
DATE / /

* Support Vector Machine.

* Advantages of SVM

1. It works well with a clear margin of separation
2. It is effective in high dimensional spaces.
3. It uses a subset of the training set in the decision function (support vectors), so it is also memory efficient.
4. Different kernel f^n's can be sepecified for the decision f^n's & its possible to specify custom kernels.

* Disadvantages of SVM.

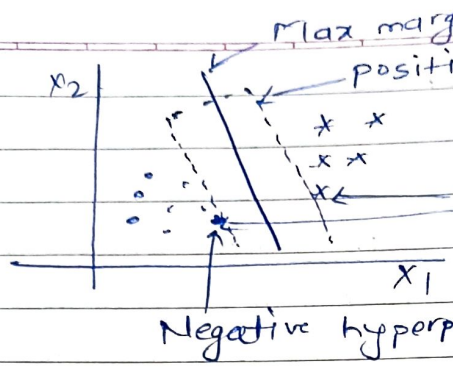
1. It does not ~~permorm~~ perform well when we have a large dataset because the required training time is higher.
2. It does not perform well when the dataset has more noise, i.e. target classes are overlapping.
3. SVM does ~~do~~ not directly provide probability estimates, these are calculated using "cross-validation.

* What is SVM.?

SVM is one of the most popular supervised ML, which is used for both classification & regression problems. However, it is ~~primaly~~ primarily, used for classification problems.

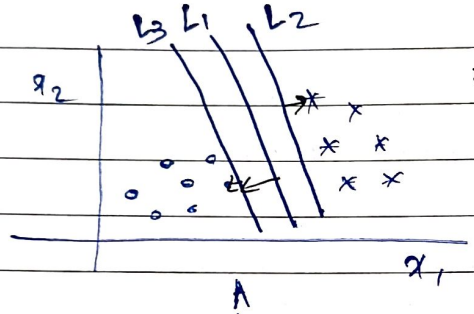
The SVM algorithm creates a best line or decision boundary that can separate n-dimensional space (n-dim features) into classes so that we can easily put the new data point in the category. This best decision boundary is called a hyperplane.

4th week. Probability & Bayes Learning.



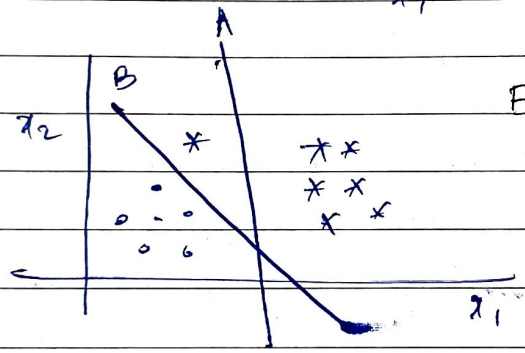
How to choose the best margine plane

(1)



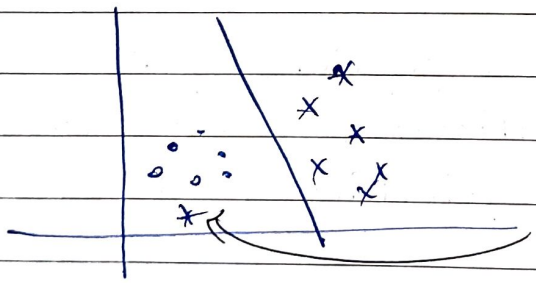
⇐ L_1 is the reasonable choice as it represents the largest margin bet^n two classes.

(2)



Even though A has larger margin than B, however SVM selects the hyper-plane B which classifies the classes accurately prior to maximizing the margin.

(3)

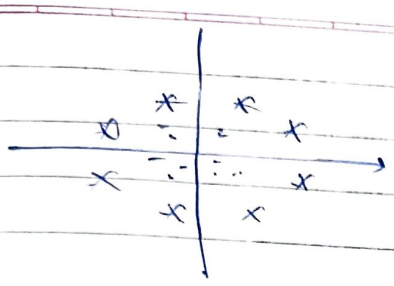


SVM algorithm has a feature to ignore outliers & find the best hyper-plane that has the max. margin.

Thus, we can say that SVM classification is robust to outliers

PAGE No.
DATE / /

④



In this case, a new feature is added $z = x^2 + y^2$.

SVM Kernel :

The SVM kernel is a function that takes low-dimensional input space & transform it into higher dimension space i.e. it converts non-separable problem to separable problem

* Hyperplane: It is a decision plane or space which divides the set of objects having different classes.

* Support vectors: The datapoints that are closest to the hyperplane are called support vectors.

* Margin: It is the gap between two lines on the closest datapoints of different classes. It can be calculated as the perpendicular distance from the line to support vectors.

* The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane in the following steps.

① First SVM will generate hyperplanes iteratively that segregates the classes in best way.

② Then, it will choose the hyperplane that separates the classes correctly.

PAGE No.
DATE / /

* Confusion matrix in ML.

This matrix is a matrix used to determine the performance of the classification models for a given ~~test~~ set of test data. It can be only be determined if the true values for the test data are known.

* Confusion matrix shows the errors in the model ~~the~~ performance in the form of a matrix, hence also known as an error matrix.

* For the 2 prediction classes of classifiers, the matrix of 2 X 2 table & so on.

*

n = no. of predictions	Acutal: Yes	Actual No
Predicted Yes	TP	FP
Predicted NO	FN	TN

True Positive TP: Model & Prediction has given Yes both

True Negative TN: -11- No.

False Positive FP: Model has predicted Yes but the actual value is No.
It is called a Type-I error.

False Negative FN: -11- Type-II error.

f1-score: F1 score is a ML evaluation metric that measures a model's accuracy. It combines the precision & recall scores of a model.

Recall: It is a metric to measure how well a ML model identifies the relevant cases in a classification task.

Precision: It is a metric to determine th how much does the model get it right when it forecasts a good outcome

PAGE No.
DATE / /

Macro avg : It represents the arithmetic mean between f1-scores of the two categories, such that both scores have the same importance.

$$= \frac{f1\_0 + f1\_1}{2}$$

* **Neural network!**

What is a neural network ?

⇒ A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain.
It is a type of ML, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. It creates an adaptive system that computers use to learn from their mistakes & improve continuously. Thus neural networks attempt to solve complicated problems, like summarizing documents or recognizing faces, with greater accuracy.

Q. Why are neural networks are important ?

⇒ Neural networks can help computers make intelligent decisions with limited human assistance. This is because they can learn & model the relationships betⁿ input & output data that are non-linear & complex. It can comphre comprehend unstructured data & make general observations without explicit training.

Q. What are neural network used for ?

→ (i) Medical diagnosis by medical image classification.

PAGE No.
DATE / /

2. Targeting marketing by social network filtering & behavioral data analysis.
3. Financial predictions by processing historical data of financial instruments.
4. Electrical load & energy demand forecasting.
5. Process & quality control.
6. Chemical compound identification.

Following are the important applications of neural network

(1) Computer vision:
Computer vision is the ability of computers to extract information from images & videos.
(i) Visual recognition in self-driving cars so they can recognize road signs & other road users.
(ii) Content moderation to automatically remove unsafe or inappropriate content from image & videos.
(iii) Facial recognition to identify faces.
(iv) Image labeling to identify brand logos, clothing, safety gear, etc.

(2) Speech recognition:
Neural networks can analyze human speech despite varying speech patterns, pitch, tone, language, & accent. Virtual assistants like Amazon Alexa sof uses speech recognition to do following tasks.
(i) Classify calls.
(ii) Convert clinical
(iii) Accurately subtitle videos &

PAGE No.
DATE / /

3. Natural language processing:
It is the process ability to process natural, human-created text. Neural networks help computers gather insights & meaning from text data & documents.
1. automated virtual agents and chatbots.
2. Automatic organization & classification of written data.
3. Business intelligence analysis of long-form documents like emails & forms.
4. Indexing of key phrases that indicates sentiment like positive and negative comments on social media.
5. Document summarization & article generation for a given topic.

4. Recommendation engines:
Neural networks can track user activity to develop personalized recommendations. They can also analyze all user behavior & discover new products or services that interest a specific user.

Q. How do neural networks work?
⇒ The human brain is the inspiration behind neural network architecture. Human brain cell, called neurons form a complex, highly interconnected network & send electrical signals to each other to help humans process information. Similarly, an artificial neural network is made of artificial neurons that work together to solve a problem. Artificial neurons are software modules, called nodes, & artificial neural networks

are software programs or algorithm that use computing systems to solve mathematical calculations.

A basic neural network has interconnected artificial neurons in three ways.

(1) Input layer:
Information from the outside world enters the artificial neural network from the input layer. Input nodes process the data, analyze or categorize it, & pass it on to the next layer.

(2) Hidden layer:
Hidden layers take their input from the input layer or other hidden layers. ANN can have large no. of hidden layers. Each hidden layer analyzes the output from the previous layer, process it further, & passes it on to the next layer.

(3) Output layer:
This layer gives the final result of all the data processing by ANN. It can have single or multiple nodes. In case of binary classification problem, the output layer will have one output node, which will give the result as 1 or 0. In case of multi-class classification problem, the output layer might consist of more than one output node.

PAGE No.
DATE / /

Deep Neural Networks:

DNN have several hidden layers with millions of artificial neurons linked together. A number, called weight, represents the connections bet^n one node to another. The weight is a positive if one node excites another, or negative if one node suppresses the other. Nodes with higher weight values have more influence on the other nodes.

The deep learning networks derives the features by itself & learns more independently. It can analyze unstructured datasets like text documents, identify which data attributes to prioritize, & solve more complex problems.