09.01.2024
Tuesday

MACHINE LEARNING → searching for the best hypotheses (starts)

Intelligence (Human intelligence)
Learn new things, creative, adapt to environment, ability to learn.

Artificial Intelligence ∴ System should replicate human intelligence

Expert System PROLOG and LISP → AI languages

Artificial Intelligence :-

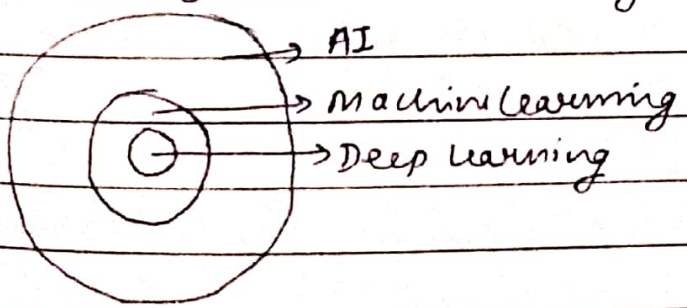
(1) Learning :- It focuses on acquiring data & creating rules for how to turn it into actionable information.

(2) Reasoning :- It focuses on choosing the right rule to reach outcome.

(3) Self correction :- It focuses to continuously fine tune rules and ensure they provide the most accurate results possible.

(4) creativity :- Generate new images, new text, new music Generative AI.

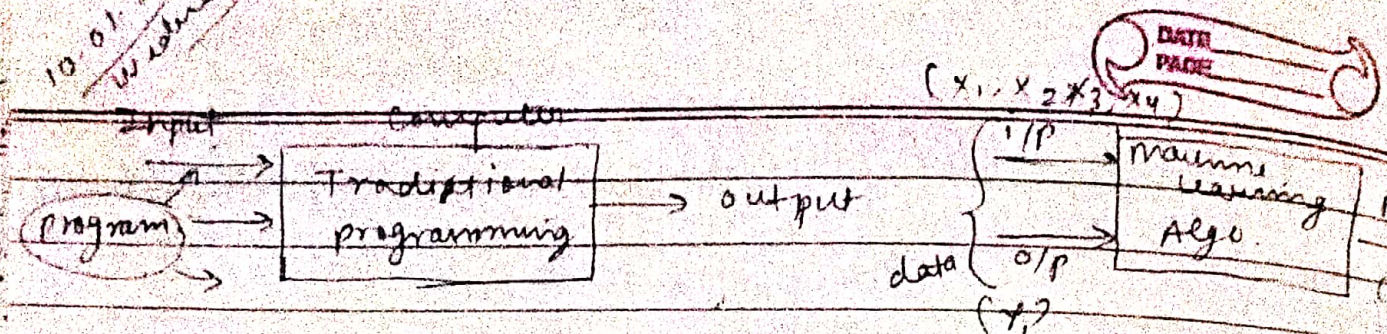
Creativity := Deep learning



Machine learning :-
Arthur Samuel (1959) "The field of study that gives computers the ability to learn without being explicitly programmed"

Scanned with CamScanner

10.01.2024
Wednesday



Supervised → I/P + O/P (data)
Unsupervised → I/P data
Ecommerce site recommendation system

Machine learning – Tom M. Mitchell
"A machine is said to learn if it is able to take experience & utilize it such that its performance improves upon similar experiences in the future."
A computer program is said to learn from experience E with respect to some task T & performance measure P, if its performance at task T, improves with experience E.

Task :- Decision Making, Prediction, Classification
Experience (Data) :- Collection of different data items related to task.
Performance :- Accuracy, precision, error.

Information related to data :-

Attributes / Features / Dimensions

A	B	C	D	E	T
					y_1
x_1	x_2	x_3	x_4	x_5	y_2
					y_3

I/P — O/P

O/P Attribute = Target attribute or O/p label or ground truth

Attributes

Rows / Tuples or instances or data points or examples

Different O/P values are called Sample space or Instance space or all possible O/P attributes

Scanned with CamScanner

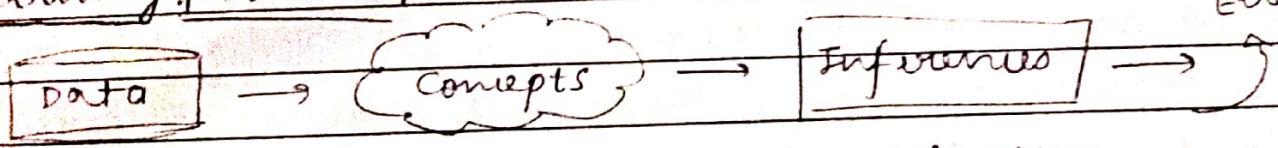
Take an example of learning process
Explain each process in detail.

Examples:-

① Handwriting recognition problem
Task:- Recognition & classification of handwritten words within images
Performance:- % of words correctly classified
Data:- Dataset of handwritten words with given classification

② A robot driving problem.
Task:- Driving on road using vision sensors
Performance:- Average distance travelled before an error.
Data:- sequence of images & steering command recorded while observing a human driver.

Learning process:-



Data Storage Abstraction Generalization

① Data:- It utilizes observation, memory storage & provide a factual basis for the further reasoning.
② Abstraction:- (i) ~~Abstracting~~ Extracting knowledge about the stored data.
③ ~~Inference~~ (ii) creating general concepts about the data as a whole
③ Generalization:- Process of turning the knowledge about stored data a form that can be utilized for future actions.
④ Evaluation:- It is the process of giving feedback to the user to measure the utility of the learned knowledge.

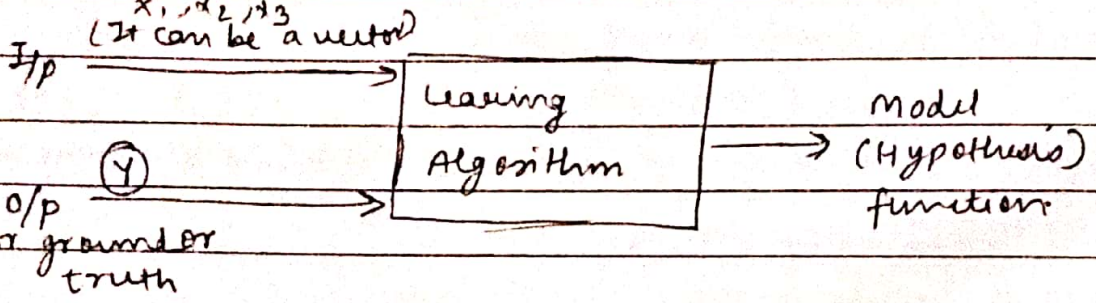
Scanned with CamScanner

15-01-2024
Monday

Types of Machine learning

1. Supervised learning
2. Unsupervised learning
3. Semi supervised learning
4. Reinforcement learning

① Supervised learning : In SL, a model is trained on a "labelled dataset" → It will have both i/p and o/p.



In SL, a training set of examples with correct responses is provided and based on this training set, the algorithm generalises to respond correctly to all possible I/p.

Types of SL

(i) Classification :- It deals with predicting of historical target variables, which represent discrete classes or labels.

Like, Yes & No., 0 and 1, Male or Female

Ex: (1) Classification of image as spam or not spam.

(2) Predicting whether a patient has high risk or low risk of heart disease

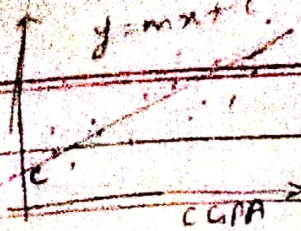
Ex: a) Decision Tree
b) Logistic Regression
c) SVM
d) Naive Bayes

(ii) Regression :- predicting continuous target variables which represents numerical values.

Regression Algo. learn to map the i/p features to a continuous numerical o/p values.

Scanned with CamScanner

Data mining – Find hidden info about data



Linear Regression

Ex: a) Linear Regression
b) Polynomial Regression

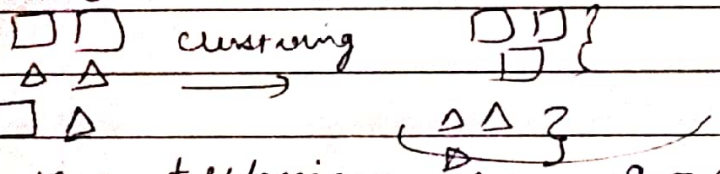
Ex: (i) Predicting the price of houses based on its size, locations, amenities.
(ii) Forecasting the sales of a product.
c) Decision Tree

(2) Unsupervised Learning :-
data is not labelled.
In unsupervised learning correct responses are not provided but instead the algorithm tries to identify similarities within the i/ps so that i/p that have something in common are categorized together.

a) Clustering
↳ grouping the data on some similarity
⇒ Kmeans Clustering

b) Association
↳ Market Basket Analysis
Bread & Butter & milk.

a) It is the process of grouping data points into cluster based on their similarities.



a.1) Kmeans algorithms
a.2) Mean-shift algorithms

b) It is a technique for discovering relationship between items in a dataset.

Ex. Market bucket analysis :- items that people buy together frequently. ex, Bread, milk, butter.

b.1) Apriori Algorithm
b.2) FD - growth Algorithm

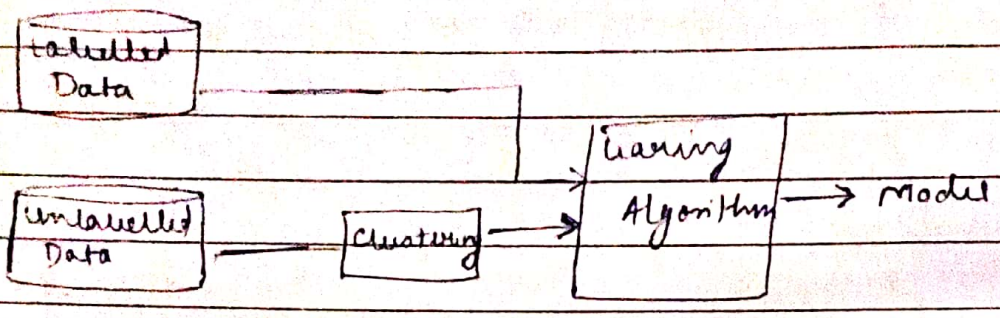
Summary we do not get labelled dataset.

(3) Semi-supervised Learning (Algorithm) :-
There are huge collection of unlabelled data & some labelled data. Labelling is a costly process & difficult to perform by the humans

Scanned with CamScanner

TPU: matrix multiplication in 1 cycle. (4×4)
Neural n/w →
hpc

The main aim of the semi-supervised learning is to efficiently use all the available data.

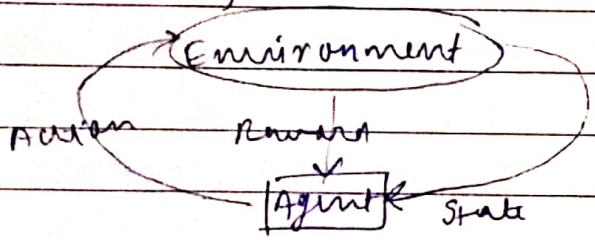


Hybrid model

Ex:- Text classification

④ Reinforcement learning :- Eg: driving cars

In it there is no I/p data as we find in supervised or unsupervised learning algo. It is about the learning the optimal behaviour or action in an environment to obtain maximum reward. A learner is not told what action. But learner will discover which action yields the most reward by trying them. (trial & error).



Algorithm
Ex: Q-learning
i) Train computer to play chess.
ii) Train the dog on new tricks

Challenges of Machine Learning :-

① Problems :- Ill for algorithm
is Task + Performance measure / Performance measures + Experience

Ill posed problem o/p

x_1, x_2	target
(1, 1)	1
(2, 1)	2
(3, 1)	3
(4, 1)	4
?	?

$y = x_1 + x_2$
$y = x_1 \cdot x_2$
$y = \frac{x_1}{x_2}$
$y = x_1^{x_2}$

Many models on same data set.

Scanned with CamScanner

17.01.2024 Wednesday

(2) Huge data :- primary requirement
↳ quantity. ~~Parameters~~
No missing data, incorrect data correct.

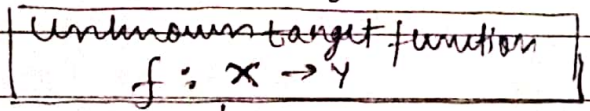
(3) High performance coding
By Data Machines with GPUs & TPU

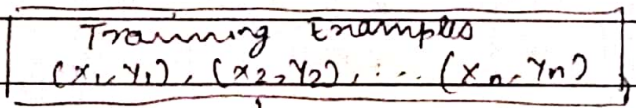
(4) Complexity of ~~appointment~~ Algorithm :-
to design
solut
evaluate

(5) Overfitting & Underfitting.
A model that too fits the traning data correctly but fails to for test data.

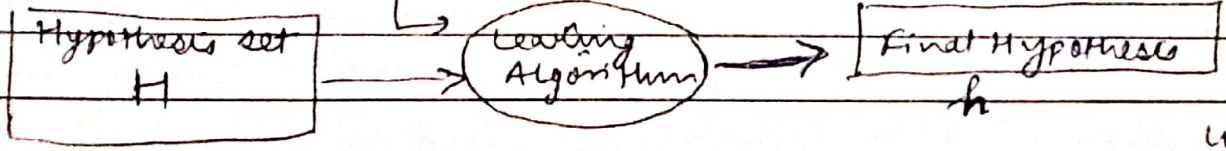
Application of Machine Learning :-
Movie Recommendation

Designing a Learning System

to 



h is subset of H



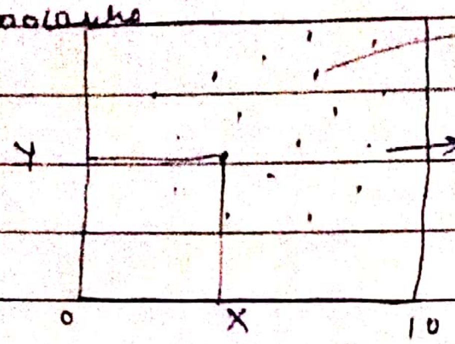
Class of Form : User decide the form → Linear → Polynomial

→ Function has 2 things { Attributes ; m & c value for linear Regression (after o/p we decide)

Scanned with CamScanner

DATE
PAGE

Feature vector :- (x_1, x_2, x_3, x_4)

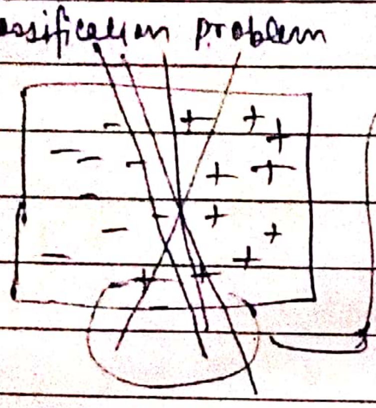


→ data points

→ feature space (All this space, infinite pts.) → I will use the subset of this.

Sample of points from this are data points.

Classification problem



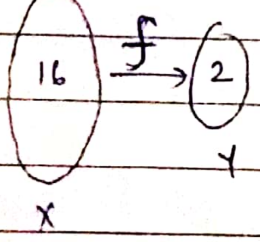
→ Hypothesis space : Set of all possible hypothesis

↳ Why hypothesis? → Because it is not proved yet

Calculation of size of hypothesis space :-

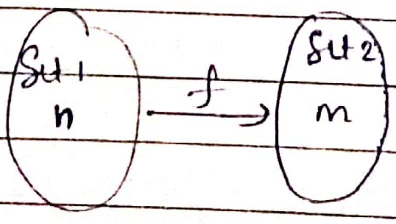
Suppose x_1, x_2, x_3, x_4 are 4 features & the value of these features is boolean. $x_1, x_2, x_3, x_4 \mid y$

~~Set 1 Set 2~~



{ Possible instances $= 2^4 = 16$ for X
{ Possible instances $= 2$ for Y

2^{16} functions are possible = Hypothesis space } → Subset after applying biases.



m & n are no. of instances

| m^n = functions possible |

One is 4 instances & other 3 instances & y has 3 instances

$3 \times 3 \times 3 \times 4 = x$ instances $= 108$

3 − Y instances 3 $\quad 3^{108}$

~~Q) If $x_1 \in R$ then~~ ~~∞~~

23.01.2024
Tuesday

DATE
PAGE

Two types of Biases :- (that can be added to Hypothesis)

① Constraints
↳ Limited Polynomial regression for Hypothesis

② Preferences
↳ I am going to select lower order polynomial

$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \ldots + w_n x^n$
(only this)

LINEAR REGRESSION

① Regression Analysis :- It is the used to model the relationship bet^n one or more independent variables x_i and independent variable y_i

Ex. prices of real estate

area (independent variable)	price (dependent variable)

(reality, city → can be other parameters that define price)

	A1	A2	A3	....	An	o/p Y
1						
2						
3						
.						
m						

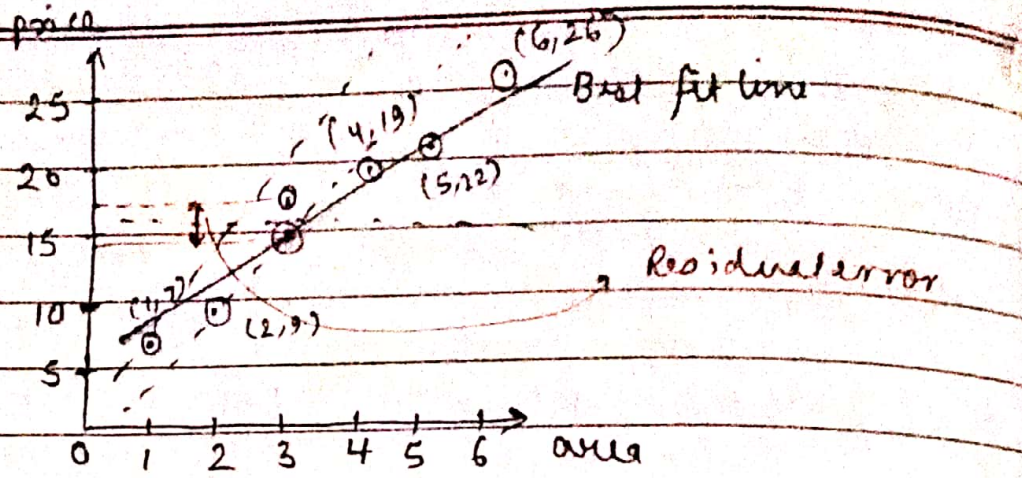
(i) No. of features are : n
(ii) No. of datapoints : m
(iii) i^{th} data point x_i $1 \le i \le m$
(iv) o/p associated with i^{th} data point : y_i
(v) predicted o/p : $\hat{y}_i$
(vi) Hypothesis function → weights
$h_w(x) = w_0 + w_1 x$
(or $y = mx + c$)

In Linear Regression :-
A line is fitted upon given training data for finding the linear ~~regression~~ relationship bet^n independent & dependent variable.

Scanned with CamScanner

Gradient Descent
L3 Basic of Neural n/w & deep learning models.

area	price
1	7
2	9
3	17
4	19
5	22
6	26



Residual error :- Difference in actual value and predicted value of $y = y_i - \hat{y}_i$

Actual error by given line = Σ Residual error = $\Sigma (y_i - \hat{y}_i)$

Linear Regression uses mean square error as loss function / cost function.

loss function / Cost function (Most used)

$$MSE = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

← Mean Square Error

Evaluation matrix
- MSE
- MAE (Mean Absolute error)
- R^2 ~~error~~ Score
- RMSE Root mean Square error

{ It's differentiation is done mostly, we put extra 1/2 in formula to make calculation easier }

$$\text{or } J = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$h_w(x) = w_0 + w_1 x$

If $w_0 = 0$ then $h_w(x) = w_1 x = \hat{y}$

$\therefore J = \frac{1}{2m} (y_i - w_1 x_i)^2$

x	y
1	1
2	2
3	3
4	4

Calculate J for
$w_1 = 0.5$
$w_1 = 1$
$w_1 = 1.5$
$w_1 = 2$

Training data demo

Scanned with CamScanner

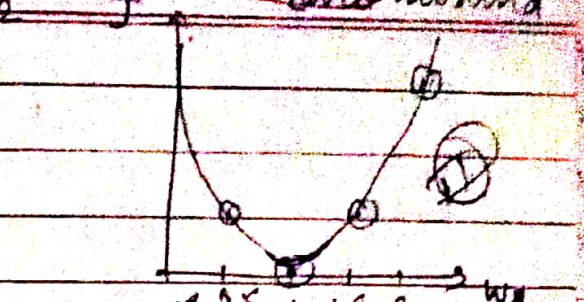
(I) Non convex function can have multiple minima but convex function has only one minima

m = 4

		$w=0.5$	$w=1$	$w=1.5$	$w=2$
		$y - wx_1$			
1	1	0.5	0	-0.5	-1
2	2	1	0	-1	-2
3	3	1.5	0	-1.5	-3
4	4	2	0	-2	-4
J		0.25	0	0.9	3.75
		2.25	7.50		
		7.50	8		
		0.9			

$J = \frac{1}{2m}(\quad)$

$\frac{30}{8} = 3.75$ $\quad \frac{16}{30}$



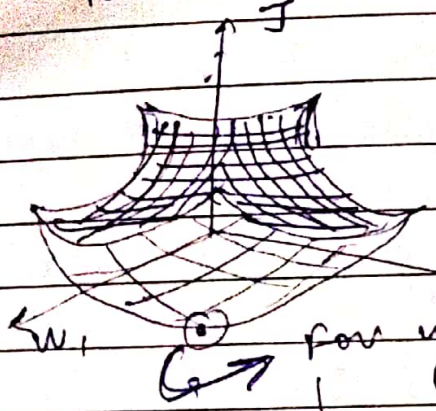
Convex function

Only 1 minima

[Same can be drawn for w_0]

$$w_{new} = w_{old} - \eta \frac{dJ}{dw}$$

When you are away then $\frac{dJ}{dw}$ is big & as you come nearer $\frac{dJ}{dw}$ becomes small.



3D convex function

for which w_0 & w_1, J value is minimum

↳ Aim

↳ Give the most optimal set for cost function

Decide Stopping Criteria → min. error permissible

→ Number of iterations

$$\text{Ideal } \frac{dJ}{dw} = 0$$

convex vs Non convex fn → (I)

Global minima vs local minima

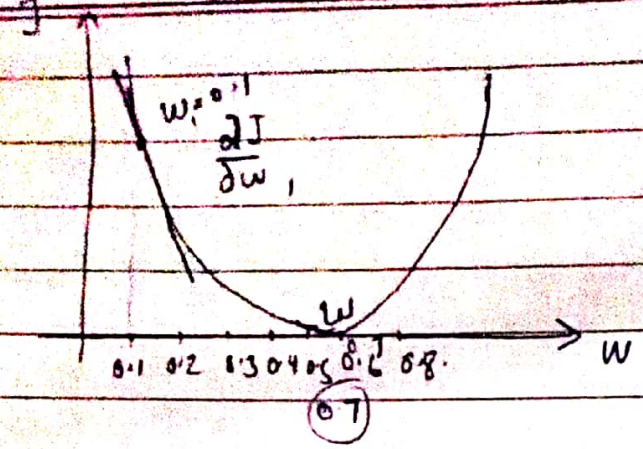
(I)

Linear Regression using Gradient Descent.

① It is an optimization technique (here minimizing)

② It is an iterative algorithm for finding a local minimum of a differentiable fn.

Scanned with CamScanner



$$w(new) = w(old) - \alpha\left(\frac{\partial J}{\partial w_1}\right)$$

For w_1 $\quad \frac{\partial J}{\partial w_1} < 0$ $\quad \left[-\frac{\partial J}{\partial w}\right]$

$$\therefore \quad w_1(n) = w_1(0) + \alpha \left|\frac{\partial J}{\partial w_1}\right|$$

Gradient Decent Algorithms (GDA):-

Step 1: choose hyperparameters (α, E) values

α = Learning rate $\quad$ E = stopping criteria

Step 2: Initialize w's values.

step 3: $w(new) = w(old) - \alpha \frac{\partial J}{\partial w}$

step 4: check stopping criteria 'E'. If satisfied stop else repeat from step (3)

$$J = \frac{1}{2m}\sum_{i=1}^{m}(y_i - h_w(x))^2 \Rightarrow J = \frac{1}{2m}\sum_{i=1}^{m}(y_i - (w_0 + w_1 x_i))^2$$

$$\frac{\partial J}{\partial w_0} = \frac{1}{2m}\sum_{i=1}^{m} 2\alpha (y_i - (w_0 + w_i x_i))^2 (-1)$$

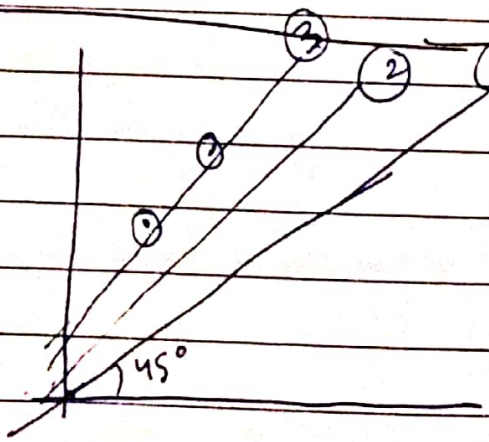
$$\Rightarrow \boxed{\frac{\partial J}{\partial w_0} = \frac{1}{m}\sum_{i=1}^{m}(w_0 + w_i x_i - y_i) \text{ or } \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)}$$

$$\boxed{\frac{\partial J}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m}(y_i - (w_0 + w_1 x_i)) \times (-w_1) \text{ or } \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i) w_1}$$

Q)

x	y
2	5
3	8

Sol:



$\hat{y} = w_0 + w_1 x$

GDA for 2 iterations

$w_0 = 0$, $w_1 = 1$

$\alpha = 0.01$

Scanned with CamScanner

Iteration-1

$\hat{y}_i$ → predicted value of y_i

$$\frac{\partial J}{\partial w_0} = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i) = \frac{1}{2}\{(2-5) + (3-8)\}$$

$$= \frac{1}{2}\{(-3) + (-5)\} = -4$$

$$\frac{\partial J}{\partial w_1} = \frac{1}{2}(-3\times 2 + (-5)\times 3) = \frac{-21}{2} = -10.5$$

$$(w_0)_n = w_0 - \alpha \frac{\partial J}{\partial w_0} = 0.04 = w_0$$

$$(w_1)_n = 0.01 \times (-10.5) = 1.105$$

Iteration-2

$$\hat{y}_i = (w_0)_n + (w_1)_n x_i \longrightarrow (2)$$

$$\hat{y}_0 = 0.04 + 1.105\, x$$

x	y	$\hat{y}$
2	5	2.25
3	8	3.355

$$\frac{\partial J}{\partial w_0} = \frac{1}{2}\{(2.25-5) + (3.355-8)\}$$

$$= \frac{1}{2}(-2.75 - 4.645) = -3.6975$$

$$\frac{\partial J}{\partial w_1} = \frac{1}{2}\{(-2.75)\times 2 + (-4.645)\times 3\} = -9.7175$$

$$w_0 = 0.04 - 0.01 \times (-3.6975) = 0.0769$$

$$w_1 = 0.04 - 0.01 \times (-9.7175) = 1.137$$

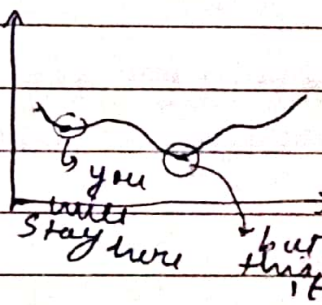
$$\hat{y} = 0.0769 + 1.137\, x \longrightarrow (3)$$

→ Line is becoming best fit

Issues in GDA :-

① Effect of loss function on GDA

In case of non-convex f^n GD can stuck on local minima



② Effect of data on GDA

	A_1	A_2	A_3	Y
0-1		1-10.5	1-10.9	

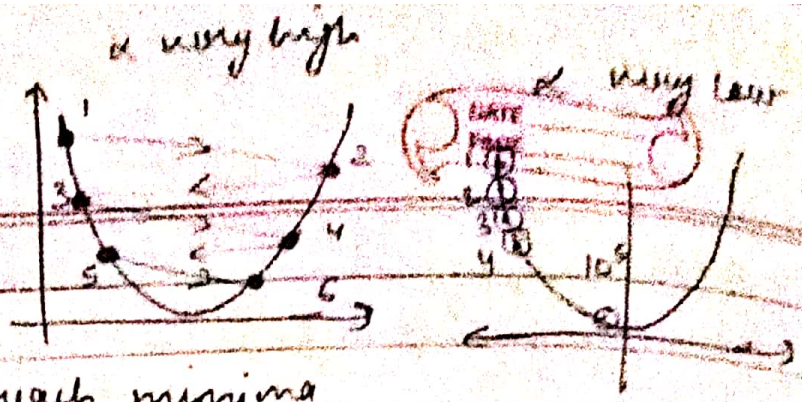
Range of attributes is different

GD will take more time to converge

∴ Scaling of values is done.

Scanned with CamScanner

29/01/2024 Monday



③ Effect of α on GDA

~~If you~~ very maybe
α is high, you will not reach minima.
α is very low, you will take much more time.
(Take various α values, generally α = 0.1 or 0.01)
Hit & trial

MULTIPLE LINEAR REGRESSION

We have multiple independent variables.

The basic assumptions of MLR:-

① Independent variables are not highly corelated. & hence ~~multicollinearity~~ multicollinearity problem does not exist.

ex: Height & weight in database. → $\sum |\hat{y}_i - y_i|$

② It is assumed that residuals are normally distributed

$x_1, x_2, Y \qquad Y = f(x_1, x_2)$

$Y = w_0 + w_1 x_1 + w_2 x_2$

For n independent variable

$Y = f(x_1, x_2, x_3, \cdots, x_n)$

$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_n x_{in}$

↳ (value of x_1 at ith row)

Linear Regression in matrix form

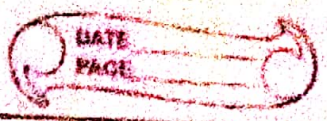
Ordinary least square approach. (OLS)

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_m \end{bmatrix}_{m \times 1} \qquad X = \begin{bmatrix} x_{11}, x_{12}, x_{13}, \cdots, x_{1n} \\ x_{21}, x_{22}, x_{23}, \cdots x_{2n} \\ \vdots \\ x_{m1}\ x_{m2}\ x_{m3} \quad x_{mn} \end{bmatrix}_{m \times (n+1)}$$

↓ w0

Scanned with CamScanner

residual = $y_i - \hat{y}_i$



$\hat{Y} = XW$

$W = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}_{(n+1)\times 1}$

RULES OF ALGEBRA

$\begin{bmatrix} a \\ b \\ c \end{bmatrix} \begin{bmatrix} a & b & c \end{bmatrix} = a^2 + b^2 + c^2$

(1) $A \cdot A^T = a^2 + b^2 + c^2 + \dots$

(2) $(AB)^T = B^T A^T$

(3) $A^T B = B^T A$

(4) $\frac{d}{dx} X^T X = 2X$

$\hat{y}$: predicted

$$J = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= (Y - XW)(Y - XW)^T \quad [\text{From } \hat{Y}]$$

$$= (Y - XW)(Y^T - (XW)^T)$$

$$= Y^T(Y - XW) - (XW)^T(Y - XW) \quad \text{* Check order}$$

$$= Y^T Y - Y^T XW - (XW)^T Y - (XW)^T XW$$

$$\Rightarrow J = Y^T Y - 2Y^T XW - W^T X^T XW \quad [A^T B = B^T A^T]$$

$$\Rightarrow \frac{dJ}{dW} = 0 - 2Y^T X - 2X^T XW = 0 \quad [\text{using } (4)]$$

$$\Rightarrow X^T XW = Y^T X$$

Multiply both $(X^T X)^{-1}$

$$\Rightarrow \boxed{W = (X^T X)^{-1}(Y^T X)}$$

or

$$\boxed{W = (X^T X)^{-1}(X^T Y)} \quad [A^T B = B^T A]$$

Q) Find Linear Regression of the data of milk & Product sales (in thousands gwen below)

Scanned with CamScanner

Closed form solution

x	y
1	1
2	3
3	4
4	8

Use LR in matrix form to find the Hypothesis

$h_w(x_i) = w_0 + w_1 x_i$

Sol:

$$y = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 8 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$X^T X = \underset{2\times4}{\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}} \underset{4\times2}{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

$$\text{adj}(X^T X) = \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix} \qquad |X^T X| = 120 - 100 = 20$$

$$(X^T X)^{-1} = \begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{bmatrix}$$

$$(X^T X^{-1}) X^T = \underset{2\times2}{\begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{bmatrix}} \underset{4\times2}{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$$

$$X^T y = \underset{2\times4}{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}} \underset{4\times1}{\begin{bmatrix} 1 \\ 3 \\ 4 \\ 8 \end{bmatrix}} = \underset{2\times1}{\begin{bmatrix} 16 \\ 51 \end{bmatrix}}$$

OLS is computationally costly operation ∴ Gradient Descent is better.

$$X^T y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 4 \\ 8 \end{bmatrix} = \begin{bmatrix} 16 \\ 51 \end{bmatrix}$$

$$(X^T X)^{-1} X^T y = \underset{2\times2}{\begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{bmatrix}} \underset{2\times1}{\begin{bmatrix} 16 \\ 51 \end{bmatrix}} = \begin{bmatrix} -1.5 \\ 2.2 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = W$$

Scanned with CamScanner

30/01/2024 Tuesday

Gradient Descent Method –

(i) Batch Gradient Descent (ii) Mini Batch GD (iii) Stochastic GD
→ subset → randomly only one data point

Regression Matrics :-

(i) mean Absolute Error (MAE) $= \frac{1}{m}\sum_{i=1}^{m} |y_i - \hat{y}_i|$ → otherwise some error + ve → cancel 0

Advantage of MAE :-

① MAE unit and y unit is same.

② It is robust to outliers. → extreme data points

Disadvantage of MAE :-

MAE Graph

① It is not differentiable at all points

(ii) Mean Squared Error (MSE) $= \frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2$

It is the sum of square of residuals.

MSE Graph

Advantage :-

① It is differentiable at all the points

Disadvantage :-

① unit of MSE and Y are not same.

② Not robust to outliers.

(iii) Root Mean Square Error (RMSE) $= \sqrt{\frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$

Advantage :-

① RMSE units is same y unit

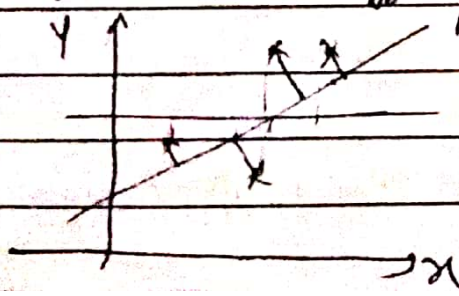
Disadvantage :-

② Not much robust to outliers.

{matrix depends on content of problem}

∴ We use R2 - Score

(iv) R2-Score :- coefficient of Determination



data not depend on x

$y = c$

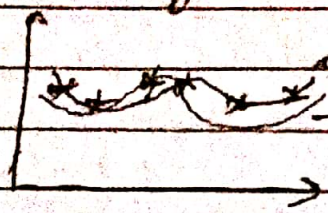
$R2\ Score = \frac{RSS}{TSS}$

RSS = Sum of squal of residual

TSS = Total sum of squares (It is sum of all squares of differences of actual value & overall mean Y.

Scanned with CamScanner

Regularization: Regularization technique used to reduce the errors by fitting the function approximately the given training set and avoiding overfitting. → appropriately

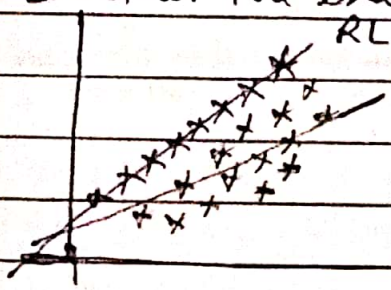


3 techniques for Regularization :- → Square / power of 2

1. Ridge Regularization (L2 Regularization) — $w \neq 0$
2. Lasso Regularization (L1 Regularization) — w can be 0
3. Elastic Net Regularization (L1 & L2 Regularization)
   ↳ = Ridge + Lasso

1. Ridge Regularization (L2 Regularization)

Consider the example -



* → testing data points
* → training data points

RL1 → zero training error + very high testing error

After regularization

Modified cost function (in OLS) (can be done by gradient descent)

$$J = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{m} w_i^2$$

↳ hyperparameter → penalty of Ridge Regularization

$$J = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda w_1^2$$

$$\hat{y}_i = w_0 + w_1 x_1$$

$$w_0 = \bar{y} - w_1 \bar{x}, \quad w_1 = \frac{\sum_{i=1}^{m} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{m} (x_i - \bar{x})^2 + \lambda}$$

$\lambda = 0 \Rightarrow$ ordinary least square $\quad [\lambda = 0 \text{ to } \infty]$

$\lambda \to \infty \Rightarrow w_1 \to 0$ but $w_1 \neq 0$

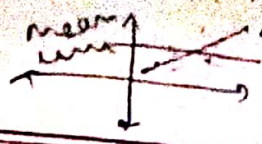
$\lambda \uparrow$ also $\Rightarrow w_1 \downarrow$ (to 0) but never becomes 0.
Reducing importance of w_1 (smoothing of curve, reducing sharpness from curve)

for $\lambda = 0$ to N : {tuning λ to get better results}
Reduce complexity & removing overfitting

Scanned with CamScanner

2 Sept 2024 Monday

$R^2 < 0$ ⇒ mean line



DATE PAGE

LASSO Regularization (Least Absolute shrinkage & selection operator)

In ridge weights will never be equals 0 which in LASSO it can be zero. (absolute function)

* $\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$ → $w_1 = 0$ (-1)

No effect of x_1 on y

* This new ⊖ Regularization.

$$J = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{m} |w_i|$$

For very high value of λ, some weights will become zero.

How some weights will become zero in Lasso?

$\hat{y}_i = w_0 + w_1 x_i$ $\quad J = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$

$w_0 = \bar{y} - w_1 \bar{x}$ $\quad w_1 = \dfrac{\sum_{i=1}^{m} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{m} (x_i - \bar{x})^2}$

$$J = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{m} |w_i|$$

↳ Not differentiable

When $w_1 > 0$

$$w_1 = \frac{\sum_{i=1}^{m} (y_i - \bar{y})(x_i - \bar{x}) - \lambda}{\sum_{i=1}^{m} (x_i - \bar{x})^2}$$

{ when $\lambda = \sum_{i=1}^{m} (y_i - \bar{y})(x_i - \bar{x})$ $\quad N^T = 0$ $\quad \therefore w_1 = 0$ }

$w_1 < 0$

$$w_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + \lambda}{\sum_{i=1}^{m} (x_i - \bar{x})^2}$$

{ Same $\lambda = \mu$ उच्चतो for $w_1 = 0$ }

Shrinking तो Normally ही | + Selection of attributes or removal of attributes

use Lasso (L1 better than L2 in this case)

Scanned with CamScanner

Dataset is too large in terms ~~of no of attributes.~~
~~If there is multicollinearity~~ (dependency)

ELASTIC NET

Elastic Net is a hybrid method of combining both Ridge and Lasso regression

Methods:

$$J = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \underbrace{\lambda_1 \sum_{i=1}^{m} w_i^2}_{\text{Ridge}} + \underbrace{\lambda_2 \sum_{i=1}^{m} |w_i|}_{\text{Lasso}}$$

When λ_1 & $\lambda_2 = 0$ both then it will work as OLS

Best $\lambda_1 = 0.5$ & $\lambda_2 = 0.5$. & then turns it

If $\lambda_1 = 0$ Then it is Lasso.

If $\lambda_2 = 0$ Then it is Ridge.

Scanned with CamScanner