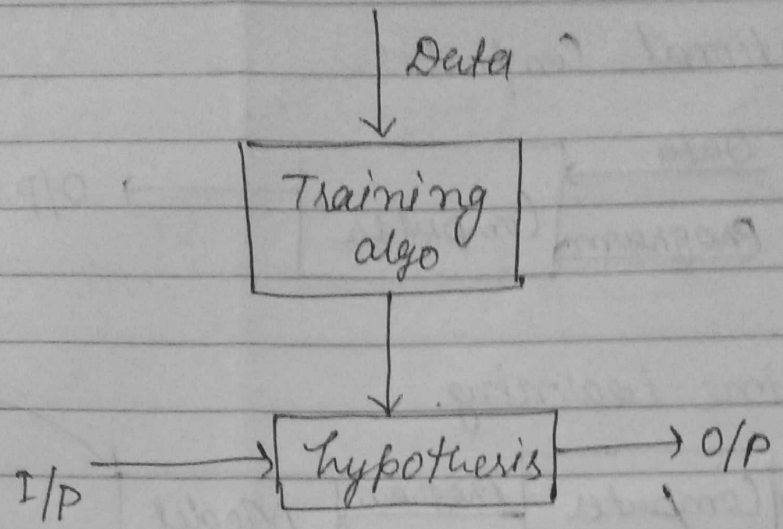
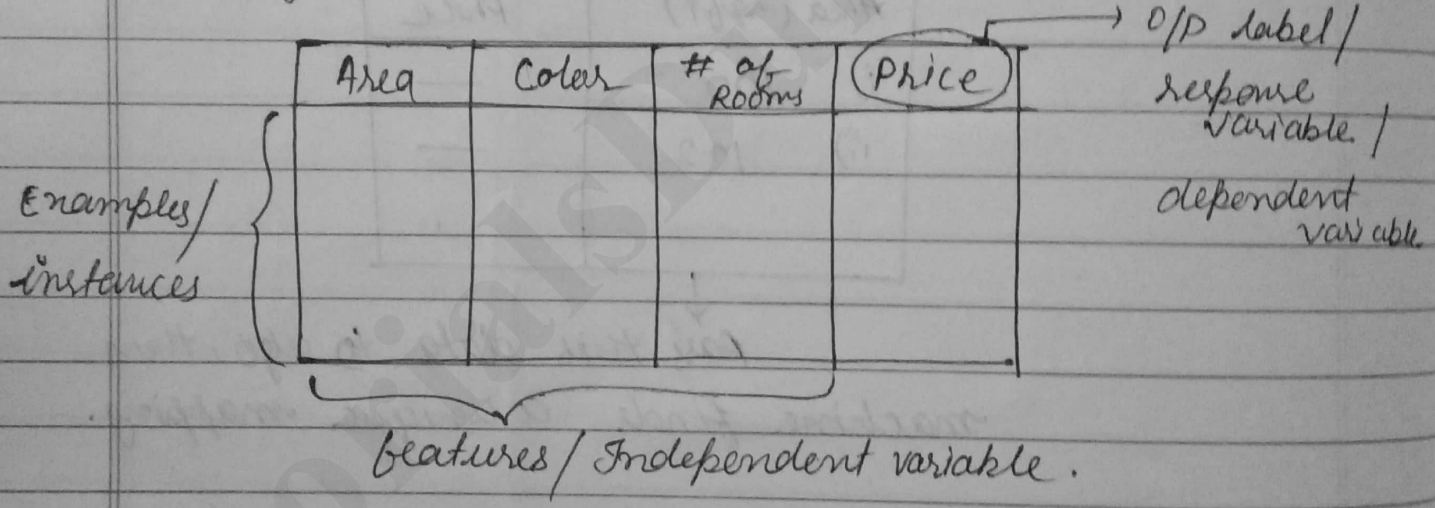
Date 3/01/19



Dataset:- A tabular structure of Data, in which data is given in form of rows & column.

e.g: in House-price detection.



Training Data 80% of total Data is use to traine the machine.

Test Data Remaining 20% of total Data is use to check the accuracy of machine

Label Data :- Data with I/P as well as O/P.

Date ___/___/_____

Types of M/L :-

i) Superwise learning.
ii) Unsuperwise learning - Clustering
iii) Reinforcement learning.

Superwise learning:-
It is the type of M/L algorithms in which training data sets consists of label data i.e. data with input values and responses value. It is also known as learning with examples. Algorithm here tries to find out the relationship between input values and output values.

Superwise Learning categorises into.
i) Regression problem
ii) Classification problem.

Data is represented as :-

- Superwise

$$D = \{(x^i, y^i)\}_1^N$$

↳ Corresponding Response.
N :- No. of Rows.

- Unsuperwise

$$D = \{x^i\}_1^N$$

e.g. Take a collection of 1000 essays written on U.S economy, and find the way to group those essays into small groups, that are some how similar or selected to each other based on feature like

Saathi

Date __/__/____

word frequency, page contentent, page count etc.
Which kind of problem it is?

:- Clustering Problem

Representation of data in Superwise learning.

$$X = \text{(input matrix)} \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(m)} \end{bmatrix}_{m \times 1} = \begin{bmatrix} x_1^1 & x_2^1 & x_3^1 & \cdots & x_n^1 \\ \vdots & & & & \\ \cdots\cdots & & x_j^i & \cdots & \\ \vdots & & & & \\ x_1^m & x_2^m & x_3^m & \cdots & x_n^m \end{bmatrix}_{m \times n}$$

$x_j^i \rightarrow$ Value of j^{th} features of i^{th} example.

$$Y = \begin{bmatrix} \vdots \\ y^i \\ \vdots \end{bmatrix}_{m \times 1}$$

y^i = outcome of i^{th} example.

In regression:

$$f(x): X \rightarrow Y$$
$$X \in R^n$$
$$Y \in R^1 \quad [\text{bcoz } Y \text{ is single dimensional}]$$

Classification:

$$f(x): X \rightarrow Y$$
$$X \in R^n$$
$$Y \in \{1 \ldots k\}$$

Saathi

Date 8, 01, 19

Application of ML :-

1) Learning Associations
   ↳ Supermarket for Basket Analysis
2) Face Recognition
3) Fradulent transactions
4) High risk / low risk
5) pattern recognition
   ↳ Handwriting recognition

Machine learning ~~wrote~~ work flow :-

1) Data collection
2) Data Exploration & Analysis
3) Model Training
4) Model evaluation
5) Model improvement

$$P(B \wedge A) = P(B/A) \cdot P(A)$$
$$P(A \cap B) = P(A/B) \cdot P(B)$$

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

Bayes Rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Theorem of total probability

if $A_1 \dots A_n$ all mutually exclusive events with

$$\sum_{i=1}^{n} P(A_i) = 1$$

Saathi

Date ___/___/_____

$$P(B) = \sum_{i=1}^{n} P(B/A_i)\ P(A_i)$$

(H)

likelihood

$$h_i \quad P(h_1/D) = \frac{P(D/h) \cdot P(h)}{P(D)}$$

Prior probability of hypo

$P(h_1/D)$ → posterior probability

$P(D)$ → evidance / probability of current data.

h_{map} = argmax aposterior prob

$$h_{MAP} = \underset{h \in H}{\text{argmax}}\ P(h/D)$$

$$h_{map} = \underset{h \in H}{\text{argument max}}\ \frac{P(D/h) \cdot P(h)}{P(D)}$$

as D become same for every $h_1, h_2 \ldots$ so drop $P(D)$

for hypothesis

$$h_{map} = \underset{h \in H}{\text{argmax}}\ P(D/h) \cdot P(h)$$

$$h_{ML} = \underset{h \in H}{\text{argmax}}\ P(D/h)$$

Lets. $P(h_1) = P(h_2) = \ldots$ so, drop $P(h)$.

Hypothesis

A Hypothesis is a certain function or rule that tries to approximate the true function (target fn') . A classifier is special case of hypothesis which is discrete value.

h_{MAP} :- Maximum A Posterior Probability

Date __/__/____

Saathi

Q. Consider a medical diagonistic problem. There are two alternative hypotesis.

1. The patient has a particular form of cancer.
2. Patient doesnt has cancer.

The lab test which is use to diagon the problem has two outcomes +ve and -ve. We have prior knowledge that over the entire population 0.008 have the disease. Furthermore the lab test is only and imperfect indicator of disease. Test returns correct positive result in only 98% of cases. and correct negative result in only 97% of the cases. For which the disease is not present. Suppose a new patient for whom lab test returns +ve should we diagonose the patient has cancer or not.

Sol: h_1 = person has cancer.

h_2 = person doesn't has cancer.

Find :-

$P(h_1/D)$?

$P(h_2/D)$?

D: Result of the lab test is +ve.

$P(\text{cancer}) = 0.008$ $\quad P(\neg\text{cancer}) = 0.992$

$P(+/\text{cancer}) = 0.98$ $\quad P(-/\neg\text{cancer}) = 0.02$

$P(-/\neg\text{cancer}) = 0.97$ $\quad P(+/\neg\text{cancer}) = 0.03$

Find:

$P(\text{cancer}/+)$

$P(\neg\text{cancer}/+)$

$h_1 = P(+/\text{cancer}) \times P(\text{cancer})$

$= 0.98 \times 0.008$

$= 0.0078$

Saathi

Date

$h_2 = (+/\text{Tcancer}) \times P(\text{Tcancer})$

$= 0.03 \times 0.992$

$= 0.0298$

Here, $h_2 > h_1$

so, $h_{MAP} = \text{Tcancer}$

Now, to change it in to hypothesis.

$$P(h_1/D) = \frac{0.0078}{0.0078 + 0.0298}$$

$$= \frac{0.0078}{0.0376} = \frac{78}{376} = \frac{39}{188}$$

$$P(h_2/D) = \frac{0.0298}{0.0078 + 0.0298}$$

$$= \frac{0.0298}{0.0376} = \frac{298}{376} = \frac{149}{188}$$

Naive Bayes

Algorithm is a ML algorithm for classification problem. It is primarily used for text classification which involves high dimensional training data. A few examples are

Spam filteration
document classification etc.

It is based on Bayes theorem and is a probilistic classified.

Saathi

Date __/__/____

→ Naive Bayes algo called Naive because it assumes that attribute values are conditionally independent given a target value.

h_{map}

↓

$$V_{map} = \underset{V_j \in V}{\operatorname{argmax}} \; P(V_j | a_1, a_2, \ldots a_n)$$

$$= \underset{V_j \in V}{\operatorname{argmax}} \; \frac{P(a_1, a_2, \ldots a_n | V_j) \cdot P(V_j)}{P(a_1, a_2, \ldots a_n)}$$

$$= \underset{V_j \in V}{\operatorname{argmax}} \; P(a_1, a_2, \ldots, a_n | V_j) \cdot P(V_j)$$

$$V_{NB} = \underset{V_j \in V}{\operatorname{argmax}} \; P(V_j) \times \prod_i^n P(a_i | V_j)$$

$$\left[P(a_1, a_2, \ldots a_n) = \prod_i^n P(a_i | V_j) \right]$$

9/01/19
Lab

esc + m → code cell to markdown // use to heading
esc + y → code cell from markdown.
Shift + entre → to execute a cell.
esc + b → new cell below any specified cell.

for heading

- to get to know whether our folder is stored.
  : pwd
  return path.

Saathi

Date __/__/____

end = ' ' → to change new line to same line with space.

print ("hello", end = ' ')

output: hello word;

- By default input fn always take string input.
- Type casting
  int (Input ("enter a number"))
- print ("Hello!", "welcome", "to", "Python")

output → Hello! welcome to python.

For separator:

print ("Hello!", "welcome", "to", sep = "\n")

Hello!
welcome
to
python

# a = 'Indians'
b = "mangoes"

print ('{0} loves {1}'.format (a, b))

output → Indians lovesmangoes

Date

Saathi

Datatype

primitive : int

advance - list, tuple

Operators

+, -, ÷, floor, division, **(exponent)

- print ("%.0.4f" %. (a/b))
  - round of till four position.

- Multiple variable assignment

  a, b, C = 10, 20, 30;

- Comment

  single ##

  multiline """ ____ """

  ''' ____ '''

- Markdow executed while comments not.
- whitespaces

- for i in range (1, 10)

  print (i) → starting index.

  1
  2
  3
  :
  9.

  → Jumping step

  for i in range (1, 10, 2)

  1
  3
  :

  for i in range (10, 1, -1).

  reverse.

Page No.

Saathi

Date function

def myfunc (score, lang, rollno);

to define a keyword.

- positional argument is placed before keyword argument.

myfunc (10, rollno = "101", ...);

- a = "Hello world"
  print (a[-1]) # → d.

- 0 1 2 3 4
  H E L L O
  -5 -4 -3 -2 -1

- len (a) :- for length of a string.

For each loop

for c in a;
  print (c.

- Slicing a: "MANGO"

print (a[0:2])
print (a[0:]) // whole
print (a[0:-2]) # MAN.
print (a[-1:]) // O

- Membership operators

print ('go' in a);
// True or false

- Split an string

a = "welcome to python".
d = a.split()
print (d)

// after splitting it convert it to list.

0 010
0 100
0 110

Saathi

Date ___/___/_____

print (type(l)) // list.
print (type(l[0])) // str.

- a.splitlines() // split at new line:
- upper() — to convert in upper case.
- Strip fun" to remove leading & trailing spaces.
- isdigit fn' check only digit.
- isalpnum : fer alphabet + number Have
- find() fun // retun index number for multiple. from left and 1st occr
- rfind()
  + running from right and return its index number.
- count() : occurance. count.
- "_".join(l);
  ↑ output : My_Python.

Join chaars.

- a.capitalize() → to capitalize 1st letter.
  a.title() : My Python.
  - capitalize every words 1st lett
- Logical operator
  2^4 0010

Page No.

TutorialsDuniya.com

Download FREE Computer Science Notes, Programs, Projects, Books PDF for any university student of BCA, MCA, B.Sc, B.Tech CSE, M.Sc, M.Tech at https://www.tutorialsduniya.com

- Algorithms Notes
- Artificial Intelligence
- Android Programming
- C & C++ Programming
- Combinatorial Optimization
- Computer Graphics
- Computer Networks
- Computer System Architecture
- DBMS & SQL Notes
- Data Analysis & Visualization
- Data Mining
- Data Science
- Data Structures
- Deep Learning
- Digital Image Processing
- Discrete Mathematics
- Information Security
- Internet Technologies
- Java Programming
- JavaScript & jQuery
- Machine Learning
- Microprocessor
- Operating System
- Operational Research
- PHP Notes
- Python Programming
- R Programming
- Software Engineering
- System Programming
- Theory of Computation
- Unix Network Programming
- Web Design & Development

Please Share these Notes with your Friends as well

Class
Date 9 / 01 / 19

Saathi

Q. Based on Naive Bayes:-

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D_1	Sunny	Hot	High	weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	weak	yes
D_4	Rain	Mild	High	weak	yes
D_5	Rain	Cool	Normal	weak	yes
D_6	Rain	Cool	Normal	strong	NO
D_7	Overcast	cool	Normal	strong	yes
D_8	Sunny	Mild	High	weak	NO
D_9	Sunny	Cool	Normal	weak	Yes
D_{10}	Rain	Mild	Normal	weak	Yes
D_{11}	Sunny	Mild	Normal	strong	Yes
D_{12}	Overcast	mild	High	Strong	yes
D_{13}	Overcast	Hot	Hot	weak	Yes
D_{14}	Rain	Mild	Mild	strong	No

Given a set of 14 training examples of the target concept play tennis where each day is described by attributes outlook, temperature, humidity wind. Use Naive Based classifies on the above data to classify the following Novel Instance

$X = \langle$ outlook = sunny, Temperature = Cool, Humidity = high, wind = strong$\rangle$

14.

Naive Bayes Expression

$$V_{NB} = \underset{V_j \in V}{\operatorname{argmax}} P(V_j) \times \prod_{i}^{n} P(a_i | V_j)$$

Page No.

Date __/__/____

Saathi

Case I V_1 = Yes.

$P(\text{Yes}) \times P(\text{outlook} = \text{sunny} | \text{yes}) \times P(\text{temperature} = \text{cool} | \text{Yes}) \times P(\text{humidity} = \text{high} | \text{yes}) \times P(\text{wind} = \text{strong} | \text{yes})$

$= \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9_3} = \frac{1}{63 \times 3} = \frac{1}{189}$

Case I V_2 = NO

$P(\text{NO}) \times P(\text{outlook} = \text{sunny} | \text{NO}) \times P(\text{temperature} = \text{cool} | \text{NO}) \times P(\text{humidity} = \text{high} | \text{NO}) \times P(\text{wind} = \text{strong} | \text{NO})$

$= \frac{5}{14} \times \frac{3}{8} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = \frac{36}{1750}$

Here, $V_2 > V_1$

So,

$V_{NB} = V_2$.

Saathi

Date __/__/____

Q.2

ID	Age	INCOME	STUDENT	CREDIT RATING	BUYS COMPUTER
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No.
3	Mid-age	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No.
7	Mid-age	Low	Yes	Excellent	Yes
8	Youth	medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Mid-age	Medium	No	Excellent	Yes
13	Mid-age	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No.

Predict the class level of a tuple X using NB classification. Given the training data. The data tuples are described by attributes age, income, student and credit rating. The class level attribut buy's computer has two distinct values yes / no. The Tuple X is as follow.

X = <age = youth, income = medium, student = yes, credit rating = fair>

Sol.

V_1 = Yes.

P(yes) X P(age = youth | yes) X P(income = medium | yes) XP(student = yes | yes) X P(credit rating = fair | yes)

$= \frac{9}{14} \times \frac{2}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9} = \frac{16}{567}$

Saathi

Date $V_2 = VNo$

$P(No) \times P$

$125 \quad 7 \quad 3 \quad 1$
875

$= \frac{5}{14} \times \frac{3}{7^5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} = \frac{6}{875}$

as, $V_1 > V_2$

$V_{NB} = V_1$

6. Consider the following six following training instances.

STATUS	FLOOR	DEPARTMENT	OFFICE SIZE	NUMBERS	RECYCLE BIN
Faculty	4	CS	medium	68	Yes
Student	4	EE	Large	112	Yes
Staff	5	CS	Medium	74	No.
Student	3	EE	small	54	Yes
Staff	4	CS	medium	86	No.
Faculty	3	EE	Large	100	Yes

① For applying Naive Bayes for statistical learning provide complete probability table for each of the predicting attributes.

② Show how a naive Bayes classifer would classfy Recycle Bin attribute for the following instance.

X = < status = student, Floor = 4, department = CS, Office size = small >

③ Find the likelyhood of recyclebin for the instance below

X = < status = faculty, floor = 2, department = CS, office size = large, members = 97 >

Page No.

Saathi

Date __/__/____

(1) Predicting table

→ For status.

STATUS	YES	NO
Faculty	2/4	0/2
Student	2/4	0/2
Staff	0/4	2/2

→ for Floor

Floor	YES	NO
3	2/4	0
4	2/4	1/2
5	0/4	1/2

→ for Department

Department	YES	NO
CS	1/4	2/2
EE	3/4	0/2

→ for Office Size.

Office Size	YES	NO
Small	1/4	0/2
Medium	1/4	2/2
Large.	2/4	0/2

2. Given instance

X = < status = student, floor = 4, department = CS, office size = small >

→ For V_1 = yes

P(yes) × P(status = student/yes) × P(floor = 4/yes) × P(department = cs/yes) × P(office size = small/yes)

$= \frac{4}{6} \times \frac{2}{4} \times \frac{2}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{96}$

→ For V_2 = No.

P(No) × P(status = student/No) × P(floor = 4/No) × P(department = cs/No) × P(office size = small/No)

$= \frac{2}{6} \times \frac{0}{2} \times \frac{1}{2} \times \frac{2}{2} \times \frac{0}{2} = \frac{1}{24}$

Page No.

Date __/__/____

Saathi

as $V_2 > V_1$

Hence $\boxed{V_{NB} = V_2}$

3. Given instance :-

X = <status = faculty, floor = 2, department = CS, officesize = large, member = 97>

16/01/19

module :- collection of pre-defined functions

```
import math
print (math.factorial(5))
```

OR

```
from math import factorial
print (factorial(5))
print (math.log (math.e))
  OP → 1.0
print (math.log (16,2)) # base = 2
                4.0
```

```
import sys
print (sys.version)
print (sys.path)
```

Page No.

Date / /

List
- as an array
- heterogeneous elements.

- List of square

```
L4 = [i*i for i in range(1,6)]
print(l4)
```

o/p → [1, 4, 9, 16, 25]

- List slicing

```
print(L4[0:4])
print(l4[-1])
```

op. [1, 4, 9, 16]
[16]

- `L.append[1,2]` # inserting list into list
- `L+=[4,5,6]` # shorthand to add no. in list
- `l.remove(3)` # delete value 3 from the list
- `del.l[2]` # delete value at index no. 2.
- `80 in l` # membership element
- To print all value

```
for i in range(len(l)):
    print(l[i])
```

Date __/__/____

Saathi

Dictionary - Collection of
Key value pair

- Key should be unique.
- Key value
Stored in random order.

```
d={}
print(type(d))
    o/p → <class 'dict'>

d={"mango":100,"apple":80}
print(d)
print(d["mango"])
```

list can also pass in dictonary.

To get all the keys.
print(d.keys())

all vals
print(d.values())

```
print(d["mango"])   # It returns error if not exist
                      in the list.
d.get("mango")      # It return NaN if it is not
                      present in list.

l1 = ["apple","papaya","guava"]
l2 = [100,200,300]
p = zip(l1,l2
```

Saatvik

Practical Date __/__/____

4 Q. Create / Define single dimension / multi - dimension arrays, and arrays with specific values like array of all ones, all zeros, array with random values within a range, or a diagonal matrix.

Concept Building

Vectors

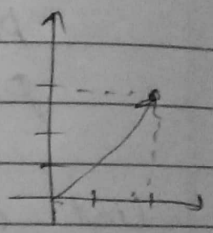
↳ magnitude + direction.

eg.

5 km/h speed → scalar

5 km/h in east → velocity.

$\begin{bmatrix} 2 \\ 3 \end{bmatrix}$ or $[2, 3] \in \mathbb{R}^2$



$\begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$ or $[3\ 2\ 1] \in \mathbb{R}^3$

$\begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \end{bmatrix} \in \mathbb{R}^n$ or $\underbrace{[1, 2, \ldots n]}_{n\text{-tuple.}}$

- Vectors are always ordered

meaning $\begin{bmatrix} 2 \\ 3 \end{bmatrix} \neq \begin{bmatrix} 3 \\ 2 \end{bmatrix}$

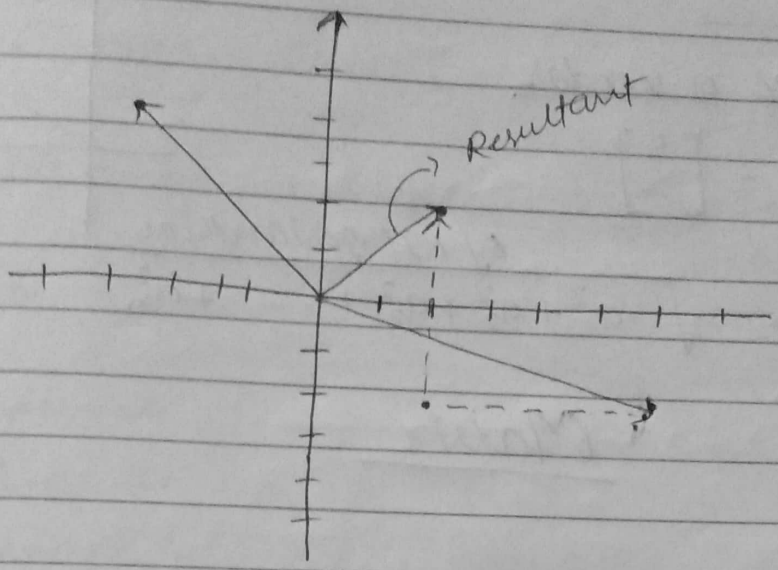
Addition of vectors :-

$\vec{a} = \begin{bmatrix} 6 \\ -2 \end{bmatrix} \quad \vec{b} = \begin{bmatrix} -4 \\ 4 \end{bmatrix}$

$\vec{a} + \vec{b} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

Date ___/___/_____

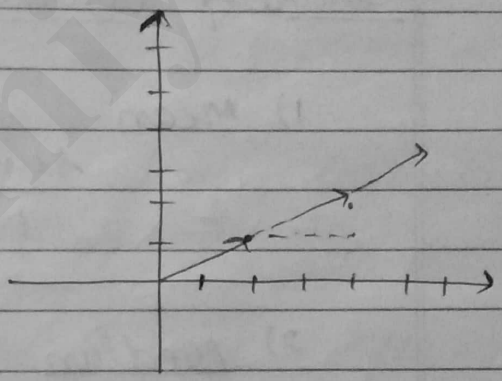
Saathi



Multiplication of vectors

$$\vec{a} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$3\vec{a} = \begin{bmatrix} 6 \\ a \end{bmatrix}$$



Linear combination of a vector

$$v_1, v_2, v_3 \dots v_n \in \mathbb{R}^n$$

$$c_1v_1 + c_2v_2 + c_3v_3 + \dots + c_nv_n$$ [linear combi

• Dot product

$$\vec{a} = \begin{bmatrix} \vec{a_1} \\ \vec{a_2} \\ \vdots \\ \vec{a_n} \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} \vec{b_1} \\ \vec{b_2} \\ \vdots \\ \vec{b_n} \end{bmatrix}$$

$$\vec{a} \cdot \vec{b} = a_1b_1 + a_2b_2 + a_3b_3 + \dots + a_nb_n$$

also:

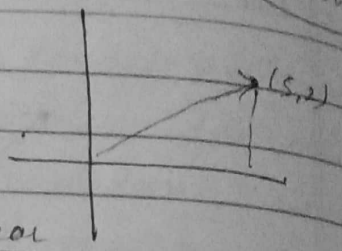
$$\vec{a} \cdot \vec{b} = a^T \cdot b = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Date ___/___/_____

Saathi

Length of a vector

$$\vec{a} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$



By pythagorian theorem

$$\|a\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \dots + a_n^2}$$

Matrix

Stat

Measure of Central Tendency

1) Mean

$$\frac{\text{sum of all observation}}{\text{No. of observation}}$$

2) Median

3) Mode (Most frequently observation).

Measure of dispersion

1) Variance

$$= \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

2) Standard Deviation

$$= \sqrt{\text{variance}}$$

$$= \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

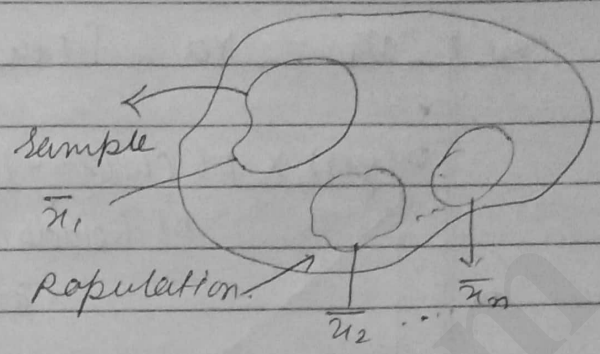
[Better bcoz S.D remains in same unit so that we can compare it easily.

Saathi

Date __/__/____

For mean (μ).

- sample mean is sample statistic which denote the sample mean



~~Sample~~ is

Statistic → represents sample

parameter → rep. population.

Sampling distribution :- Distribution of sample statistics.

17/01/19

Q. Given all prev patter patients I have seen, below are their symptoms and diagonis.

	Chills	running nose	headache	fever	flu
1.	Y	N	Mild	Y	N
2.	Y	Y	No	N	Y
3.	Y	N	Strong	Y	Y
4.	N	Y	Mild	Y	Y
5.	N	N	No	N	N
6.	N	Y	Strong	Y	Y
7.	N	Y	Strong	N	N
8.	Y	Y	Mild	Y	Y

Classify if the patients with following symptoms has flu or not.

X = <chills = Y, running nose = N, headache = Mild, fever = Y, flue = ?>

$P(h/D) = \frac{[P(D/h) \times P(h)]}{P(D)}$

Saathi

Date ___/___/_____

Case I V_1 = Yes [Flu is yes]

P(yes) × P(Chills = y / yes) × P(runnynose = N / yes) × P(headache = Mild / yes) × P(fever = y / yes)

$= \frac{5}{8} \times \frac{3}{8} \times \frac{1}{5} \times \frac{2}{5} \times \frac{4}{5} = \frac{3}{125}$

Case 2 V_2 = NO

P(NO) × P(Chills = y / NO) × P(runnynose = N / NO) × P(headache = Mild / NO) × P(fever = y / NO)

$= \frac{3}{4} \times \frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{108}$

$V_1 > V_2$

so, $\boxed{V_{NB} = V_1}$

Hence it classify that patient shows the symptom has flu.

$P(V_1) = \frac{3/125}{\frac{3}{125} + \frac{1}{108}} = \frac{3}{125} \times \frac{125 \times 108}{108 \times 3 + 125} = \frac{324}{324 + 125}$

$= \frac{324}{449}$

$P(V_2) = \frac{1/108}{\frac{3}{125} + \frac{1}{108}} = \frac{1}{108} \times \frac{125 \times 108}{3 \times 108 + 125} = \frac{125}{449}$

Page No.

UNIT -3

Date 22/01/19

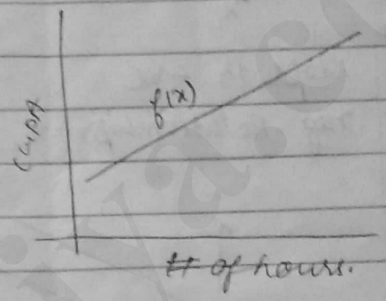
Linear Regression

- O/P variable is continous then regression problem.
- O/P variable is discrete then classification problem.

Under Regression problem, we plot the data.

# of hours	CGPA
2	6
4	7
6	7.5
10	8.5

(7)



as f(x) is line so $f(x) = mx + c$.

↑ we have to find

→ Linear regression is a linear approach for modeling the relationship between a dependent variable and one or more explanatory variables. When we have single input or explanatory variable then the linear regression is known as simple linear Regression. whereas linear regression incorporating multiple input variables is known as multiple linear regression.

→ we are trying to find a linear relation b/w input and output variable.

Our goal is to find eq. like this → (Best possible line).

$y = m(x) + c$

- target / response variable. ↑ dependent variable (every case only one)
- ← Independent variable / explanatory variable / Predictor variable (may be one or more)
  - ↓ simple
  - ↳ multiple.

TutorialsDuniya.com

Download FREE Computer Science Notes, Programs, Projects, Books PDF for any university student of BCA, MCA, B.Sc, B.Tech CSE, M.Sc, M.Tech at https://www.tutorialsduniya.com

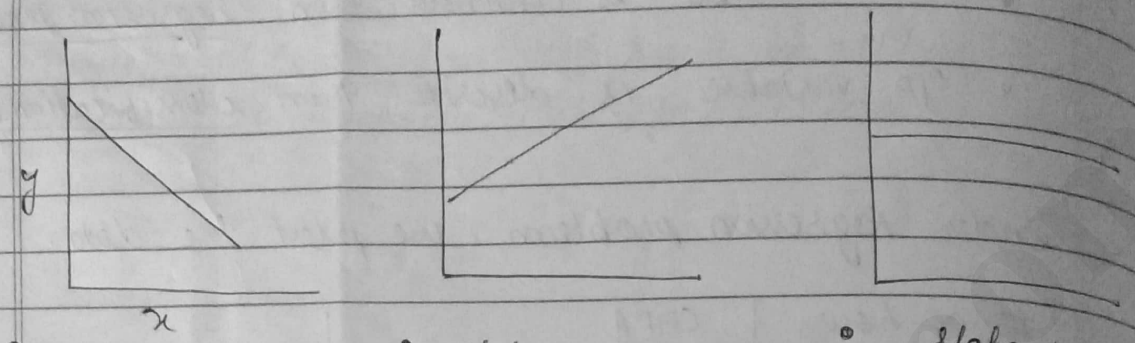
- Algorithms Notes
- Artificial Intelligence
- Android Programming
- C & C++ Programming
- Combinatorial Optimization
- Computer Graphics
- Computer Networks
- Computer System Architecture
- DBMS & SQL Notes
- Data Analysis & Visualization
- Data Mining
- Data Science
- Data Structures
- Deep Learning
- Digital Image Processing
- Discrete Mathematics
- Information Security
- Internet Technologies
- Java Programming
- JavaScript & jQuery
- Machine Learning
- Microprocessor
- Operating System
- Operational Research
- PHP Notes
- Python Programming
- R Programming
- Software Engineering
- System Programming
- Theory of Computation
- Unix Network Programming
- Web Design & Development

Please Share these Notes with your Friends as well

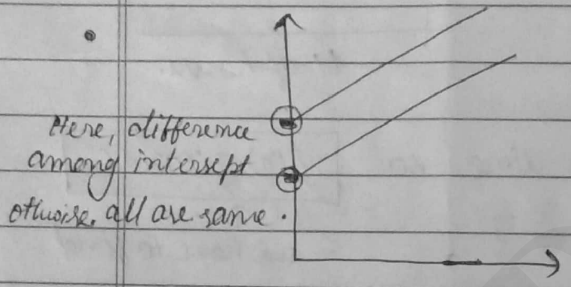
Saathi

Date __/__/____

Examples of Linear Relationship:-



- Slope is -ve
  neg. Relationship.
- Slope is +ve
  +ve Relation
- Slope is 0
  constant Relation



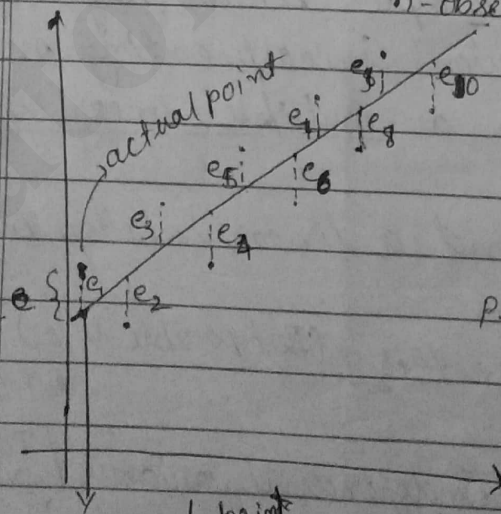
$y = c + mx$

$y = \beta_0 + \beta_1 x$

Goal:- To find β_0 & β_1 such that resulting line is close to true line.

data $(x_1, y_1)\ (x_2, y_2)\ \dots\ (x_n, y_n)$

n-Observations



Predicted line

$\hat{y}_i = \beta_0 + \beta_1 x_i$

actual line

Predicted line $\hat{y}_i - y = e_i$

Residual sum of square

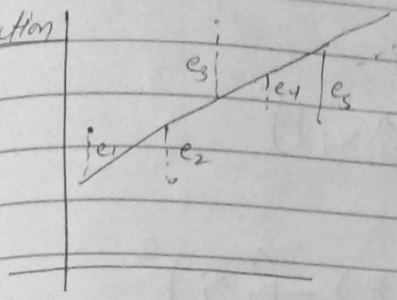
$= e_1^2 + e_2^2 + \dots + e_n^2$

Goal: to minimise RSS.

Ist Jan syllabus
Date / /

Saathi

Derivation



$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ least square line

$e_i = y_i - \hat{y}_i$ ← residual of the ith example

$$RSS = e_1^2 + e_2^2 + \dots + e_n^2$$
$$= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2) + \dots + (y_n - \hat{y}_n)^2$$
$$= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2) + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

To find minimum.

diff w.r.to β_0 :-

$$\frac{\partial RSS}{\partial \beta_0} \Rightarrow \sum_{i=1}^{n} (-)2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$ ← to calculate minima

$$\Rightarrow \sum_{i=1}^{n} y_i - \hat{\beta}_0 \times n - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i = n\hat{\beta}_0$$

Divide by n both side

$$\Rightarrow \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \frac{n\hat{\beta}_0}{n}$$

$$\Rightarrow \boxed{\bar{y} - \hat{\beta}_1 \bar{x} = \hat{\beta}_0}$$

as, $\frac{\sum_{i=1}^{n} y_i}{n}$ - average of y_i

$\frac{\sum_{i=1}^{n} x_i}{n}$: avg of x_i

Page No.

Date ___/___/___

diff w.r.t β_1.

$$\frac{\partial RSS}{\partial \beta_1} \Rightarrow \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i\right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i\right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})\right)^2$$

Now, diff.

$$\frac{\partial RSS}{\partial \beta_1} \Rightarrow \sum_{i=1}^{n} -2 (x_i - \bar{x}) \left(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \sum_{i=1}^{n} \beta_1 (x_i - \bar{x})^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\Rightarrow \beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Q. Given the data $D = \{(1,2), (2,3), (4,6)\}$ predict a value for $x = 3$. Assume that a simple linear model

$\bar{x} = \frac{1+2+4/3}{3} = 7/3$ $\quad \bar{y} = 1/3$

Date ___/___/_____ Saathi

i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1.	$-4/3$	$-5/3$	$16/9$	$20/9$
2.	$-1/3$	$-2/3$	$1/9$	$2/9$
3.	$5/3$	$7/3$	$25/9$	$35/9$

$$\beta_1 = \frac{57/9}{42/9} = 57/42 = 1.357$$

$$\beta_0 = \frac{11}{3} - \frac{57}{42} \times \frac{7}{3} = \frac{22-19}{6} = 3/6 = 1/2$$

Anal. exp. $\hat{y}_i = \beta_0 + \beta_1 \hat{x}_i$

$y_3 = 1/2 + \frac{57}{14} \times 8$ Here, $x_i = 3$

$= \frac{7+57}{14} = \frac{64}{14} 32 = \frac{32}{7} = 4.57$

Q. Find the least square line for the following data points. (1, 1) (2, 3) (4, 3), (3, 2) (5, 5).

$\bar{x} = \frac{1+2+4+3+5}{5} = 3$ $\quad \bar{y} = \frac{1+3+3+2+5}{5} = \frac{14}{5}$

i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	-2	$-9/5$	4	$18/5$
2	-1	$1/5$	1	$-1/5$
3	1	$1/5$	1	$1/5$
4	0	$-4/5$	0	0
5	2	$11/5$	4	$22/5$
			Total = $\frac{40}{5} = 8$.	

$\beta_1 = \frac{40/5}{10} = \frac{8}{10} = 0.8$ $\quad \beta_0 = 14/5 - 0.8 \times 3$

$= \frac{8}{10} - 2.4 = 14/5 - \frac{24}{10} = \frac{4}{10} = 0.4$

$8/10 \times 3$

$\hat{y}_i = 0.4 + 0.8 \hat{x}_i$

Date ___/___/___

RSS

i	y_i	$\hat{y}_i$
1	1	1.2
2	3	2
3.	3	3.6
4	2	2.8
5.	5	4.4

$e_1 + e_2 + e_3 + e_4 + e_5$

$= (-.2) + (1) + (-0.6) + (-0.8) + (.6)$

RSS = $e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$

$= 0.04 + 1 + 0.36 + 0.64 + 0.36$

$= 2.4$

RSS = 2.4

Another formular for slope (B_1) :-

$$B_1 = \frac{\text{co-veriance}(x, y)}{\text{veriance}(x)}$$

we are regressing y on x

where, equation of line is for.

$y = \beta_0 + \beta^2 x$

↑ x is the predicting value

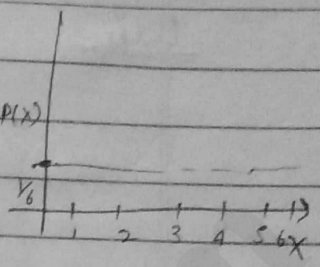
variance of input value is as denominator

Date 23, 01, 19

Saathi

Rolling an unbiased dice:-

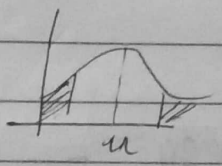
	1	2	3	4	5	6
P	1/6	1/6	1/6	- - -		1/6



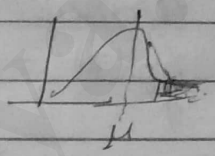
Uniform Distribution (continuous)

$\bar{x} = u$ } 2 tailed test
$\bar{x} \neq u$

$\bar{x} > u$ } 1 tailed test
$\bar{x} < u$



$h_0 \neq B_1 = 0$
$h_{a1} B_1 \neq 0$



① Confidence Interval approach

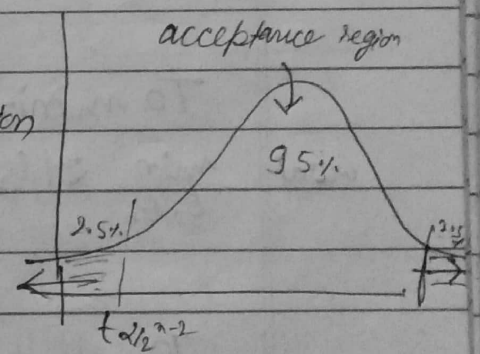
Sample estimate ± t-multiplier × SE

$\alpha = 95\%$

dependson
- α
- degree of freedom

② Rejection region approach:-

→ If 2 tailed test - 2 sided Rejection Region.
so leftover region is divided in two parts.



→ If 1 tailed test - 1 side rejection.

shaded - rejection region
Non-shaded - acceptance region

→ If h_0 Null hypothesis is true so, it lies on Reject region

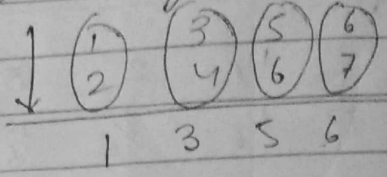
Probility value

* Less the p-value, evidence against Null hypothesis is more significance.

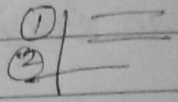
Saathi

Date 30/01/19

Lab min (a, axis=0) along x-axis



if axis 1 so along y-axis



~~31/01/18~~ 31/01/19

In another terminology :- estimated slope

$$h_\theta(x) = \theta_0 + \theta_1 x \quad \text{——— (1)}$$

$h_\theta(x)$ ↑ hypothesis. θ_0 ↓ est. Intercept

Average error

$$J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)^2$$

($h_\theta(x^i)$: Hypothesis value; y^i: original value)

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x^i - y^i\right)^2$$

$J(\theta_0, \theta_1)$

To minimize the above we add $\frac{1}{2}$ in above

Goal $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

$$= \frac{1}{2m}\sum\left[\theta_0 + \theta_1 x^i - y^i\right]^2$$

To simplify ~~eq~~ (1)

$$h_\theta(x) = \theta_1 x \quad [\text{intercept} = 0]$$

so, $J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)^2$

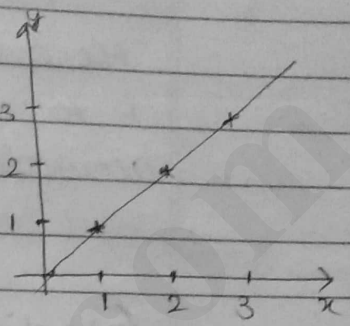
$$= \frac{1}{2m}\sum_{i=1}^{m}\left(\theta_1 x^i - y^i\right)^2$$

Date ___/___/_____

Saathi

Objective min $J(\theta_1)$

e.g. data points (1,1), (2,2), (3,3)

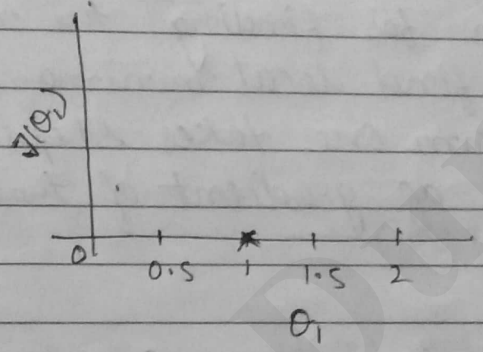


$J(\theta_1 = 1)$

$= \frac{1}{2\times3}(0^2+0^2+0^2)$

$= 0$

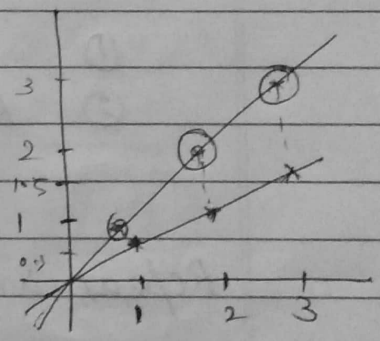
Graph of cost function



$J(\theta_1 = 0.5)$ $[y = 0.5 \times x]$

$= \frac{1}{2\times3}[(0.5)^2 + 1^2 + (1.5)^2]$

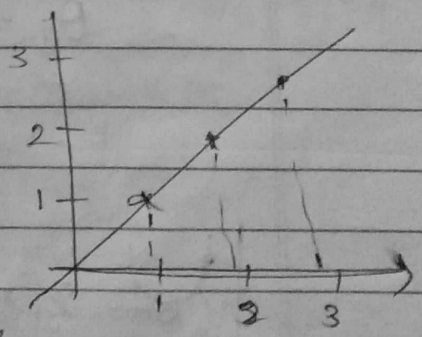
$= 0.58$



$J(\theta_1 = 0)$ $[y = 0]$

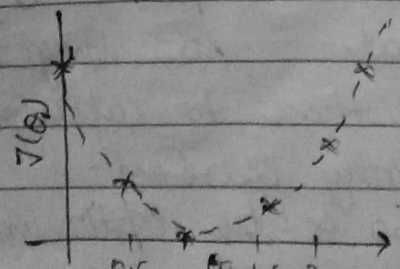
$= \frac{1}{2\times3}[1^2+2^2+3^2]$

$= 7/3 = 2.3$



Hence for different θ_1 value we observe that it is convex fun. having single local minima.

*** For min value of $J(\theta_1)$ we have more actual line.

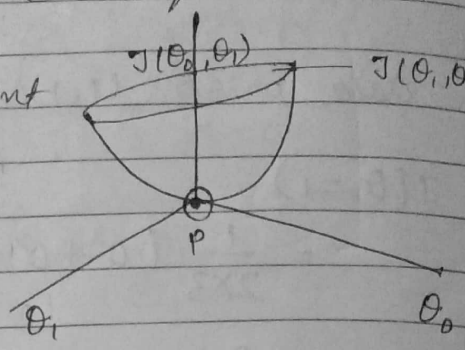


Page No.

Saathi

Date __/__/____

Now For $J(\theta_0, \theta_1)$ the cost function is.



Here we need to find point P to attain the more accurate line.

Gradient decent :-

↳ slope ↳ dec.

Gradient decent is a first order iterative optimization algorithm for finding the minimum of a function. To find local minima of a function using this algorithm one takes steps proportional to the negative of gradient of that function at that point.

① Start from any θ_0, θ_1

② keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$ until we converge.

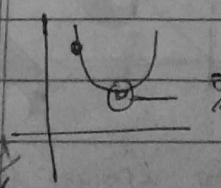
Repeat until convergence

$$\theta_j = \theta_j - \alpha \left| \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \right|$$

learning rate.

slope of this curve at

Here we moving in -ve direction of slope.



so at this point $\theta_j = \theta_j$

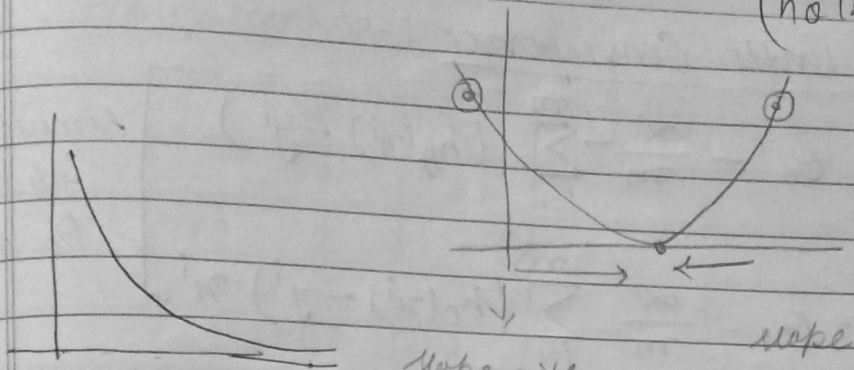
If α is greater so longer step and algo take very less time to calculate but for more large it may overshoot.

So α be optimum.

If α is very less then it takes too long to

Date 1/02/19

clear [X, Y, theta] (h_θ(x) - y^i)^2 Saathi



slope -ve

slope +ve
to θ1 decreased and move forward

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum (h_\theta(x^i) - y^i)^2 \qquad h_\theta(x) = \theta_0 + \theta_1 x^i$$

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = ? \quad \text{(diff w.r.t } \theta_0\text{)}$$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum (\theta_0 + \theta_1 x^i - y^i)^2$$

[(1+0-0)

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{2m} \times 2 \sum (\theta_0 + \theta_1 x^i - y^i) \times 1$$

$$\boxed{\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum (h_\theta(x^i) - y^i)}$$

differentiate w.r. to θ_1.

$\theta_0 + \theta_1 x^i$ → $h_\theta(x^i)$ theta

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{2m} \times 2 \sum (\theta_0 + \theta_1 x^i - y^i) \times x^i$$

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum (h_\theta(x^i) - y^i) \times x^i$$

grad = [,]

Page No.

Saathi

Date

Repeat until Convergence :-

{

$$\theta_0 = \theta_0 - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)$$

$$\theta_1 = \theta_1 - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)\,x^i$$

} simultaneously update θ_0 & θ_1

Simultaneous update

$$\theta_0 = \theta_0 - \alpha\left[\frac{\partial}{\partial\theta_0} J(\theta_0,\theta_1)\right] \rightarrow \text{tempo}$$

$$\theta_1 = \theta_1 - \alpha\left[\frac{\partial}{\partial\theta_0} J(\theta_0,\theta_1)\right] \text{ store in temp1.}$$

then update the whole value by

$\theta_0 = \theta_0 - \text{tempo}$

$\theta_1 = \theta_1 - \text{temp1}$ so that we dont have updated value for θ_1.

Q. Find the least square regression model using gradient descent method for following dataset keeping alpha : 0.1

Carry out process for two iterations.

X	1	2	4	3	5
Y	1	3	3	2	5

Formula

$$\theta_0 = \theta_0 - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i) \rightarrow \text{for } i^{th} \text{ example}$$

$$\theta_1 = \theta_1 - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)\cdot x^i$$

Page No.

Date ___/___/___

Saathi

First Iteratation

x	Y	$h_\theta(x^i) = \theta_0 + \theta_1 x$
1	1	0
2	3	0
4	3	0
3	2	0
5	5	0

θ_0

For i

$$\theta_0 = 0 - \frac{0.1}{5}\left[(-1)+(-3)+(-3)+(-2)+(-5)\right]$$

$$= 0 - \frac{0.1}{5}\left[-14\right]$$

$$= \frac{1.4}{5}$$

$$\theta_1 = 0 - \frac{0.1}{5}\left[(-1)\times 1 + (-3)\times 2 + (-3)\times 4 + (-2)\times 3 + (-5)\times 5\right]$$

$$= 0 - \frac{0.1}{5}\left[-1 - 6 - 12 - 6 - 25\right]$$

$$= \frac{0.1}{5} \times 50 = 1$$

Second Iteration

$\theta_0 = \frac{7}{25}$, $\theta_1 = 1$

~~$\theta_0 = \frac{7}{25} - 0$~~

x	Y	$h_\theta(x^i) = \theta_0 + \theta_1 x$
1	1	$7/25 + 1 = \frac{32}{25}$
2	3	$7/25 + 2 = 57/25$
4	3	$7/25 + 4$ $107/25$
3	2	$7/25 + 3$ $82/25$
5	5	$7/25 + 5 = 132/25$

Page No.

64
14
78

Date ___/___/_____

Saathi

$$\theta_0 = \frac{7}{25} - \frac{0.1}{5}\left[\frac{7}{25} - \frac{18}{25} + \frac{32}{25} + \frac{32}{25} + \frac{7}{25}\right]$$

$$= \frac{7}{25} - \frac{0.1}{5} \times \frac{70^{14}}{25}$$

$$= \frac{7 - 1.4}{25} = \frac{5.6}{25} = 0.232$$

$$= 0.23$$

$$\theta_1 = 1 - \frac{0.1}{5}\left[\frac{7}{25} \times 1 - \frac{18}{25} \times 2 + \frac{32}{25} \times 4 + \frac{32}{25} \times 2 + \frac{7}{25} \times 5\right]$$

$$= 1 - \frac{0.1}{5}\left[\frac{42}{25} + \frac{224}{25} - \frac{36}{25}\right]$$

$$= 1 - \frac{0.1}{5} \times \frac{266 - 36}{25}$$

$$= 1 - \frac{0.1}{5} \times \frac{230}{25}$$

$$= \frac{125 - 23}{125} = \frac{102}{125} = 0.816$$

Here it mentioned for two iteration so we stop here. otherwise

$J(\theta) <$ threshold value.

to stop for this $J(\theta)$.

Date ___/___/_____

Saathi

Solve with analytical method:-

$$\theta_1 = \beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\theta_0 = \beta_0 = \bar{y} - \theta_1 \bar{x}$$

(1, 1) (2, 3) (4, 3) (3, 2), (5, 5)

$$\bar{x} = \frac{1+2+4+3+5}{5} = \frac{15}{5} = 3$$

$$\bar{y} = \frac{1+3+3+2+5}{5} = \frac{14}{5}$$

i	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	1	1	-2	$-9/5$	4	$18/5$
2	2	3	-1	$+1/5$	1	$-1/5$
3	4	3	1	$1/5$	1	$1/5$
4	3	2	0	$-4/5$	0	0
5	5	5	2	$11/5$	4	$22/5$
					10	$40/5 = 8$

$$\beta_1 = \frac{8}{10} = 0.8$$

$$\beta_0 = \frac{14}{5} - 0.8 \times 3$$

$$= \frac{14}{5} - \frac{24}{10} = \frac{140 - 120}{50} = \frac{20}{50} = \frac{2}{5}$$

$$= 0.4$$

TutorialsDuniya.com

Download FREE Computer Science Notes, Programs, Projects, Books PDF for any university student of BCA, MCA, B.Sc, B.Tech CSE, M.Sc, M.Tech at https://www.tutorialsduniya.com

- Algorithms Notes
- Artificial Intelligence
- Android Programming
- C & C++ Programming
- Combinatorial Optimization
- Computer Graphics
- Computer Networks
- Computer System Architecture
- DBMS & SQL Notes
- Data Analysis & Visualization
- Data Mining
- Data Science
- Data Structures
- Deep Learning
- Digital Image Processing
- Discrete Mathematics
- Information Security
- Internet Technologies
- Java Programming
- JavaScript & jQuery
- Machine Learning
- Microprocessor
- Operating System
- Operational Research
- PHP Notes
- Python Programming
- R Programming
- Software Engineering
- System Programming
- Theory of Computation
- Unix Network Programming
- Web Design & Development

Please Share these Notes with your Friends as well

Date

Saathi

:- Multiple Regression :-

feature				
x_1	x_2	x_3	x_4	y → Response
2014	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178

x^i — i^{th} example

x^i_j ← j^{th} feature of i^{th} example.

$$x^2 = \begin{bmatrix} 1416 \\ 3 \\ 2 \\ 40 \end{bmatrix} \qquad x^2_2 = 3$$

- m is no. of example
- n number of feature.

$$h_\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

if we incorporate $x_0 = 1$

$$h_\theta = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

it may used as.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

Page No.

Date __/__/____

Saathi

after incorporating x_0 we get

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_m \end{bmatrix} \quad \text{and} \quad \theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_m \end{bmatrix}$$

so $\boxed{h_\theta = \theta^T x}$ → Transpose of θ.

$$\boxed{J(\theta_0, \theta_1, \ldots \theta_m) = \frac{1}{2m} \sum (h_\theta(x^i) - y^i)^2}$$

Updated Gradient descent formulae:-

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1 \ldots \theta_m)$$

so,

$$\theta_0 = \theta_0 - \frac{\alpha}{m} (h_\theta(x^i) - y^i) \cdot x_0^i \quad [x_0^i = 1]$$

and

$$\theta_1 = \theta_1 - \frac{\alpha}{m} (h_\theta(x^i) - y^i) x_j^i$$

So genral updated formula is

$$\boxed{\theta_j = \theta_j - \frac{\alpha}{m} (h_\theta(x^i) - y^i) x_j^i}$$

and $\boxed{x_0^i = 1}$ for all x_0^i.

Page No.

Date 5 / 02 / 19

Normal Equation Method

x_0 remains always 1, so feature starts from x_1

	x_0	x_1	x_2	x_3	x_n	y
m	1	-	-	-	-	-
	1	-	-	-	-	-
	1	-	-	-	-	-
	1	-	-	-	-	-
	1					

X y

- n :- features
- m = examples.

So, From Data Set.

$[X]_{m \times (n+1)}$ → as $n+1$ feature including x_0

$[Y]_{m \times 1}$

Normal Equation is :-

$$\theta = (X^T X)^{-1} X^T y$$

$(n+1) \times m$ $\quad$ $m \times (n+1)$

$(n+1)(n+1)$ $\quad$ $(n+1) \times m$

$(n+1) \times m$ $\quad$ $m \times 1$

$(n+1) \times 1$

Date / /

Saathi

Gradient descent	Normal Equation
1. Iterative	1. Non - iterative
2. Can't work with large value of n	2. for large value of n it doesn't work.
	3. has scaling feature

1. Scaling Feature.
2. Mean Normalization of feature $\left[\frac{\text{value} - \text{mean}}{\text{range}}\right]$

Q. Normalize the given data using mean-Normalization
12, 34, 45, 15, 40, 34.

$$\text{Mean} = \frac{12+34+45+15+40+34}{6} = \frac{180}{6} = 30$$

Range = 45 - 12 = 33,

$$\frac{12-30}{33}, \frac{34-30}{33}, \frac{45-30}{33}, \frac{15-30}{33}, \frac{40-30}{33}, \frac{34-30}{33}$$

$$= \frac{-18}{33}, \frac{4}{33}, \frac{15}{33}, \frac{-15}{33}, \frac{10}{33}, \frac{4}{33}$$

$$= -0.545, 0.121, 0.454, -0.454, 0.30, 0.121$$

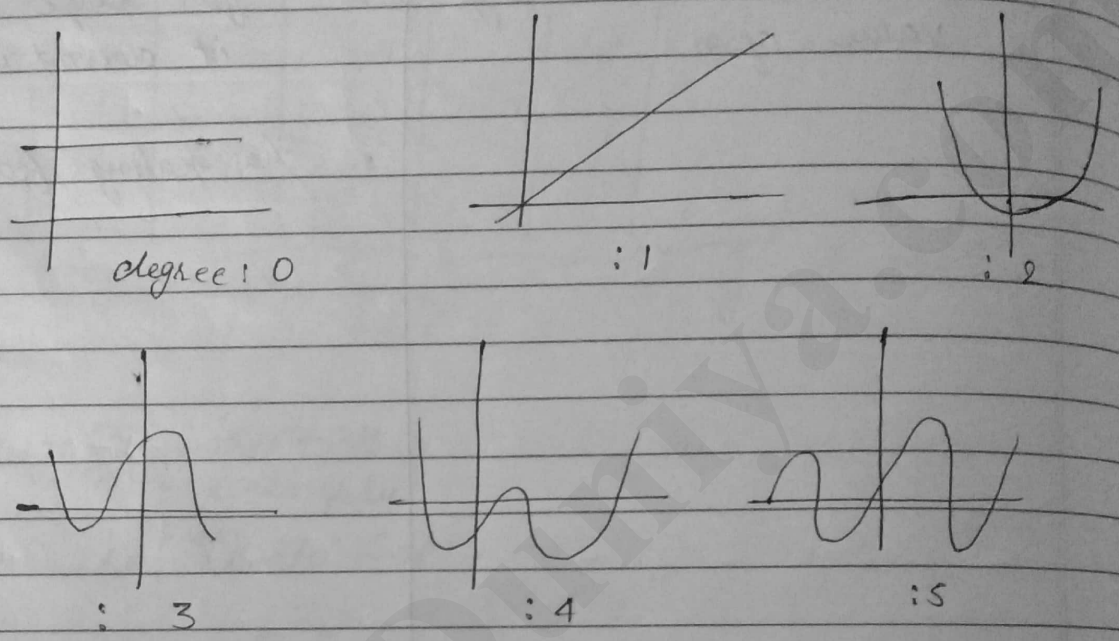
Page No.

CH 6 :- Page 203-209

Date

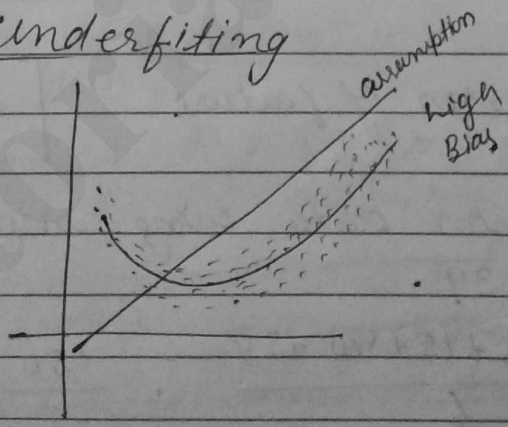
Model :- is a system that maps input to output.

degree of polynomimal



Degree represents flexibility of the model ie it can accomodate more points.

Underfiting



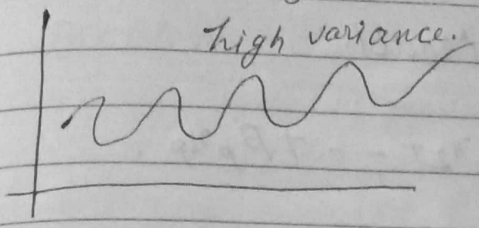
If we assume that it is linear. So we represent it as a linear expression. So our model is not able to find the relationship

Hence, this is underfitting with high Bias (high error)

On both case. Training data as well as Test data.

Date / Overfitting

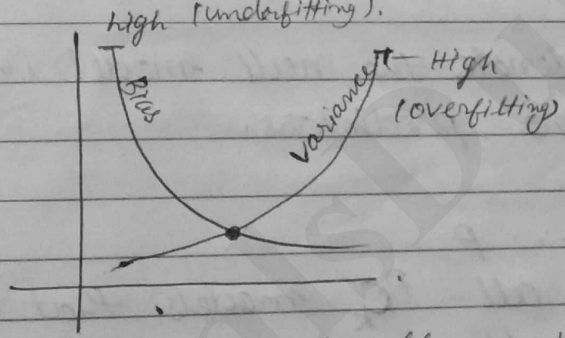
Saathi



- It lost its predictable capapibility.
- It generate error during test data. because it have more flexibility

Bias :- Bias occurs when an algorithm has limited flexibility to learn true relationship from a data set.

Variance It referss to an algorithm sensitivity to specific set of training data. It is the amount of error that the estimate of target function will change if different training data was use.

→ 

- Bias High → variance low
- Variance high → Bias low.

Bias - Variance Trade off is the problem of simultaneously minimizing two sources of error that prevents superwise learning algorithm from generalizing beyound training set.

12/02/19

prediction Accuracy :-

n → # of examples
p → # of feature.

- True linear → low Bias
  - low variance
  - high variance
- n >> p
- n > p
- n < p
  - Infinite variance

Model Interpretability.

Page No.

Date ___/___/_____

Saathi

Standard Linear Model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Least Square Approach :-

3 Methods to improve the performance of Least Square Approach of model to the data

1. Subset Selection
2. Shrinkage
3. Dimensionality Reduction.

Algo for Subset Selection :-

Algo-1 Best Subset Selection

1. Let M_0 denote the null model, which contains no predictors.

2. for $k = 1 \ldots p$
   (a) Fit all ${}^{p}C_k$ models that contains exactly k predictors.
   (b) pick the best among these ${}^{p}C_k$ models and call it M_k. Best is defined in terms of smallest RSS or largest R^2.

3. Select a single best model from $M_0 \ldots M_p$ using cross validated prediction error, AIC, BIC or adjusted R^2.

- R^2 (Coefficient of determination)
  Higher value or close to 1 is more accurate
  range of R^2 is $\boxed{0 < R^2 < 1}$

Page No.

Date ___/___/___

Saathi

Problem :

1.) Computational problem. (Computationally infeasible)

$P_{C0} + P_{C1} + P_{C2} + \cdots + P_{CP}$

$= 2^P$.

Higher Computation (which is not easy to find also for computer).

2.

Algo-2 For Backward Subset Selection (widely used)

1.) Let M_0 denote the null model, which contains no predictors.

2) for $K = 0, \cdots, p-1$

(a) Consider all $p-k$ models that augment the predictors in M_k with one additional parameter.

(b) Choose the best among $p-k$ models and call it M_{k+1}. Here best is defined in terms of smallest RSS or largest R^2.

3) Select a single best model from among $M_0 \cdots M_p$ using cross validated prediction error, AIC, BIC or adjust R^2.

	# of features	# of model
M_0 {}	1	p-0
M_1 {1}	2	p-1
M_2 {1,2}	3	p-2
:	:	:
M_{p-1} {1, -- p-1}	p	1
M_p		

Page No.

as algo-1, algo-2 (subset selection work for least square - so it won't work for n<p. approximation.

Saathi

Date ___/___/_____

Problem is P_1 and all next is dependent on this. But it may be possible that $P_2 P_3 P_4$ is best model.

But it resolve the computational issue (Algo-1) for. $p = 20$

Algo-1, $2^{20} = 1048576$

Algo-2 $1 + \frac{p(p+1)}{2} = 1 + \frac{20 \times 21}{2}$

$= 211$.

which is far less than.

Algo-3 Backward Subset Selection

1.) Let M_0 denote the null model, which contains no. predictors.

2) for $k = p, p-1, \ldots, 1$.

a) Consider all k models that contains all but one of the predictors in M_k for total of k-1 predictors.

(b) Choose the best among these k models and call it M_{k-1}. Here best is defined as having smallest RSS or highest R^2.

3) Select single best Model $M_0 \ldots M_p \ldots$

Date __/__/____

Saathi

	# of feature	# of models
M_p	{1 -- p}	p
M_{p-1}	{1 -- p-1}	p-1
⋮		
M_1	{1}	1
M_0	0	

as, for $n < p$ least square approach can't work. so, Here in Backward Subset selection it rejected at its first approach. But in forward selection it work until $n < p$. then rejected.

13/02/19 (Lab).

Q. Load and explore data from a .csv file using pandas, get its first 10 rows, print shape of data.

Q. Perform house price prediction using Boston dataset in Scikit learn package. Divide data into training and test data (80%, 20%). Calculate accuracy, plot data and predicted line.

Hint) from sklearn.datasets import load_boston
from sklearn linear_model import LinearRegression.

```
boston = load_boston()
print(boston.data)        — x feature
print(boston.target).     — y set.
```

Saathi

Date __/__/____

1. From pandas import *
df = read_csv (r"C:\Users ").
df

df.shape

df.head(10).

19/02/19. Unit-4 Logistic Regression

- Qualitative Response :-
  Categorical response, discrete response
  (when Response have categories)

- Quantitative Response :-
  Continous Response

Logistic Regression : is not a Regression algo
it is a classification algo.
while Linear Regression is a Regression algo.

Classification algo
- Linear Discriminant Analysis
- K-nearest neighbour.

- Dummy Variable approach (in 2-variable)
  Not possible for 3 attribute.

imp. Q. Why we (can't) ~~have to~~ use linear regression for Quantitative response variable.
describe with 3 point (Reasons).

In Logistic Regression we want probability as output.

Page No.

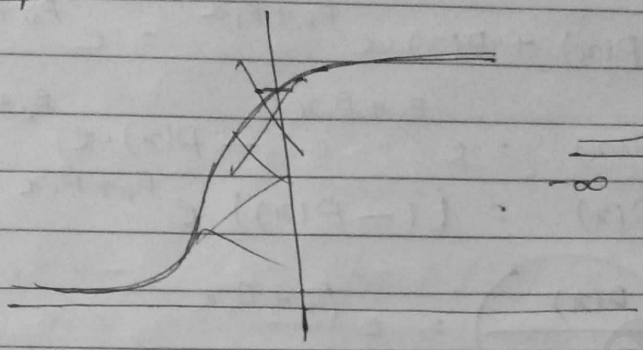
Date __/__/____

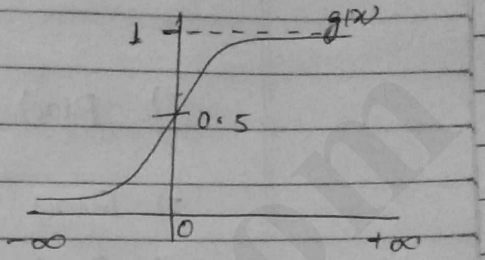
Saathi

Logistic function (sigmoid function). $g(x) = \frac{1}{1+e^{-x}}$

output : [0, 1]

graph





$P(x) = \beta_0 + \beta_1 x$ it gives output as any number (+ve, -ve) so we map res with help of logistic function to set output b/w [0,1]. for value less than 0 it outputs 0 for value greater than 0 it gives output greater than 0.5.

$$g(x) = \frac{1}{1+e^{-x}} = \frac{1}{1+\frac{1}{e^x}} = \frac{e^x}{1+e^x}$$

In Linear Regression we have

$$h_\theta(x) = \beta_0 + \beta_1 x$$

so, Probability is $P(x) = \beta_0 + \beta_1 x$ it may go beyond 0-1 range so we apply logistic function.

and Logistic Regression.

Probability of (P(x)) belonging to class

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

Estimation for β_0 & β_1 :

- Maximum-likelihood Estimation.

Page No.

Saathi

Date ___/___/_____

#

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

$$\Rightarrow P(x) + P(x) \cdot e^{\beta_0 + \beta_1 x} = e^{\beta_0 + \beta_1 x}$$

$$\Rightarrow P(x) = e^{\beta_0 + \beta_1 x} - P(x) \cdot e^{\beta_0 + \beta_1 x}$$

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

Odds ←

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

Taking log both side.

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

↳ log of odds / logit

If we increase x by 1 unit then log of odds increases β_1.

• Here logit is linear in nature.

Table ①

	Coefficient	Std. error	Z-statistic	P value
Intercept	-10.615	0.3612	-29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

②

response is default = yes

	Coefficient	Std. error	Z-statistic	P-value
Intercept	-35.041	0.0707	-49.55	<0.0001
student[yes] balance	0.4049	0.1150	3.52	0.0004

Page No.

Date __/__/__

Saathi

For Balance $1000. (From Table ①)

and $$P(x) = \frac{e^{B_0 + B_1 x}}{1 + e^{B_0 + B_1 x}}$$

$$= \frac{e^{-10.615 + 0.0055 \times 1000}}{1 + e^{-10.615 + 0.0055 \times 1000}}$$

$$= 0.00576 < 0.5$$

Which is less than 1%. So not a defaulter.

If $P(x)$ 0.5 then 50% belongingness to defaulter category.

Q. For Table ②

$P(\text{default} = \text{yes} / \text{student} = \text{yes})$ as student = yes so $x = 1$

$$P = \frac{e^{-3.5041 + 0.4049 \times 1}}{1 + e^{-3.5041 + 0.4049 \times 1}}$$

If we have to find:

$P(\text{default} = \text{yes} / \text{student} = \text{No})$

So, Here $x = 0$.

$$P = \frac{e^{-3.5041 + 0.4049 \times 0}}{1 + e^{-3.5041 + 0.4049 \times 0}}$$

$$= \frac{e^{-3.5041}}{1 + e^{-3.5041}}$$

TutorialsDuniya.com

Download FREE Computer Science Notes, Programs, Projects, Books PDF for any university student of BCA, MCA, B.Sc, B.Tech CSE, M.Sc, M.Tech at https://www.tutorialsduniya.com

- Algorithms Notes
- Artificial Intelligence
- Android Programming
- C & C++ Programming
- Combinatorial Optimization
- Computer Graphics
- Computer Networks
- Computer System Architecture
- DBMS & SQL Notes
- Data Analysis & Visualization
- Data Mining
- Data Science
- Data Structures
- Deep Learning
- Digital Image Processing
- Discrete Mathematics
- Information Security
- Internet Technologies
- Java Programming
- JavaScript & jQuery
- Machine Learning
- Microprocessor
- Operating System
- Operational Research
- PHP Notes
- Python Programming
- R Programming
- Software Engineering
- System Programming
- Theory of Computation
- Unix Network Programming
- Web Design & Development

Please Share these Notes with your Friends as well

Lab

Date 20/02/19.

Q. Implement linear regression using gradient descent algorithm:-

1) randomly pick theta's.
2) repeat until convergence
update Theta's.

$$\theta_0 = \theta_0 - \eta \sum_{i=1}^{m} (h_\theta(x^i) - y^i)$$

$$\theta_1 = \theta_1 - \eta \sum_{i=1}^{m} (h_\theta(x^i) - y^i) x^i$$

21/02/19

Simple logistic

Table

	Coefficient	Std. error	Z-Statistic	P-Value
Intercept	3.5041	0.0707	-49.55	< 0.0001
Student[Yes]	0.4049	0.1150	3.52	0.0004

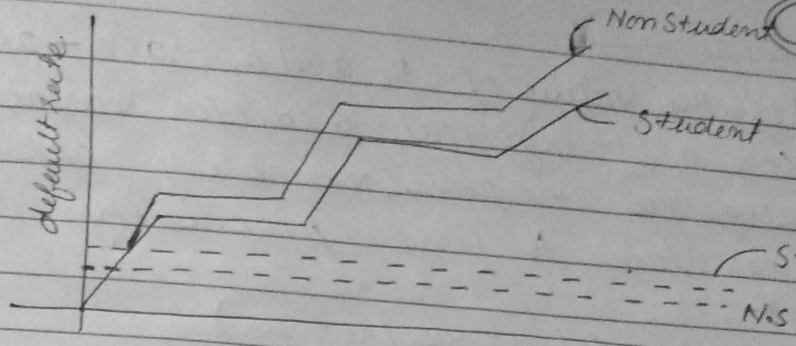
$$\log\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

$$\Rightarrow P(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

Multiple logistic

	Coeffb	Std. error	Z-Statistic	P-Value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.34	0.7115
Student[Yes]	-0.6468	0.2362	-2.74	0.0062

Page No.



---- (average default rate)

When we not consider the other aspect then we find that average default rate of student is high as compared to non-student. But in case of multiple logistic we find that Student is less defaulter than Non-Student.

Q. Predict the probability of default for student and non-student with a credit card balance $1500 and income $40,000.

• P(default / yes, student = yes).

$$p = \frac{e^{-10.8690 + 0.0057 \times 1500 + 0.0030 \times 40 + (-0.6468) \times 1}}{1 + e^{-10.8690 + 0.0057 \times 1500 + 0.0030 \times 40 + -0.6468 \times 1}}$$

$$= \frac{e^{-10.8690 + 8.55 + 0.012 - 0.6468}}{1 + e^{-10.8690 + 8.55 + 0.012 - 0.6468}}$$

$$= \frac{e^{-2.9538}}{1 + e^{-2.9538}} = \frac{0.0355}{1 - 0.0355} = \frac{0.0355}{0.9645}$$

$$= \frac{0.052}{1 + 0.052} = \frac{52}{1052}$$

Saathi

Date __/__/____

Q. For the value of B_0 and B_1 as -2.16 and 0.425 for a categorical predictor variable x and a categorical response variable y apply logistic regression to find probability of

$P(y = yes / x = yes)$

& $P(y = yes / x = no)$

$x = 1$

$$P_1 = \frac{e^{-2.16 + 0.425}}{1 + e^{-2.16 + 0.425}} = \frac{e^{-1.735}}{1 + e^{-1.735}} = \frac{0.176}{1 + 0.176} = 0.149$$

$$= \frac{0.983}{1 + 0.98} = \frac{98.49}{98.99} = 0.49 = \frac{49}{99}$$

$$P_2 = \frac{e^{-2.16}}{1 + e^{-2.16}} = \frac{0.55}{1 + 0.55} = \frac{11}{31} = 0.35$$

$$= \frac{0.115}{1 + 0.115} = \frac{0.115}{1.115} = 0.103$$

22/02/19 **Logistic Regression**

goal → $0 \leq h_\theta(x) \leq 1$

where $h_\theta(x) = \theta^T X$ in case of linear Regression

Here $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$ $X = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$

For Logistic Regression.

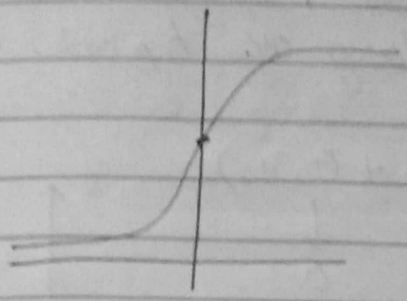
$$h_\theta(x) = g(\theta^T x)$$

Where $g(z) = \frac{1}{1 + e^{-z}}$

Date ___/___/___

Saathi

ie. $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$



$\theta^T x \geq 0 \quad [P=1]$

$\theta^T x < 0 \quad [P=0]$

- $h_\theta(x) = 0.7$ [belongs to class 1]

Probability of belonging to that class (class 1) will.

$P(y=1/x;\theta) = 0.7$

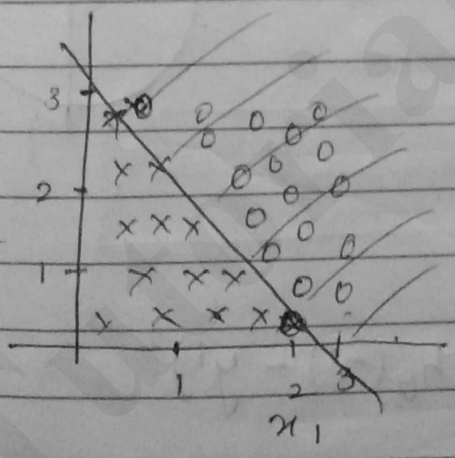
class 1 because $h_\theta(x) > P(y=0/x;\theta)$ ie $0.7 > 0.3$.

\# $P(y=1/x;\theta) + P(y=0/x;\theta) = 1$

$P(y=1/x;\theta) = h_\theta(x)$

$P(y=0/x;\theta) = 1 - h_\theta(x)$.

Decision Boundary



$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

Suppose : $-3 \quad 1 \quad 1$

$-3 + x_1 + x_2 > 0$

$x_1 + x_2 > 3$

Here we got a decision line which bounded the solution set. But in case of linear regression we haver only one line which represents point.

Page No.

Saathi

Date __/__/____

Q. Consider logistic regression with two features x_1 and x_2. Suppose $\theta_0 = 5$, $\theta_1 = -1$ and $\theta_2 = 0$. Show the decision boundary of $h_\theta(x)$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

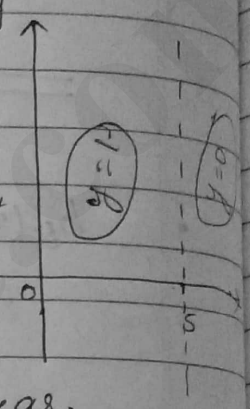
$$\downarrow \quad \downarrow \quad \downarrow$$

$$5 \quad -1 \quad 0$$

$$h_\theta(x) = g(5 - x_1)$$

$$5 - x_1 \geq 0 \quad [\text{For positive input bcoz only +ve input we have outcome as greater than 0.5}]$$

$$5 \geq x_1$$



Decision boundary can be non-linear.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2$$

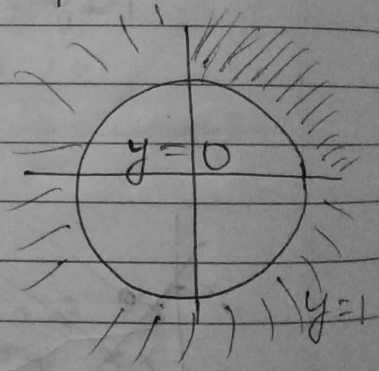
$$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$

$$-1 \quad 0 \quad 0 \quad 1 \quad 1$$

$$\theta^T x > 0$$

$$-1 + x_1^2 + x_2^2 > 0$$

$$x_1^2 + x_2^2 > 1$$



Cost Function

$$J(\theta) = \frac{1}{2m} \sum_{i=0}^{m} (h_\theta(x^i) - y^i)^2$$

$$= \frac{1}{2} \sum_{i=0}^{m} \frac{1}{m} (h_\theta(x^i) - y^i)^2$$

$$\text{cost}(h_\theta(x^i), y^i)$$

$$= \frac{1}{2} \sum_{i=0}^{m} \text{cost}(h_\theta(x^i), y^i)$$

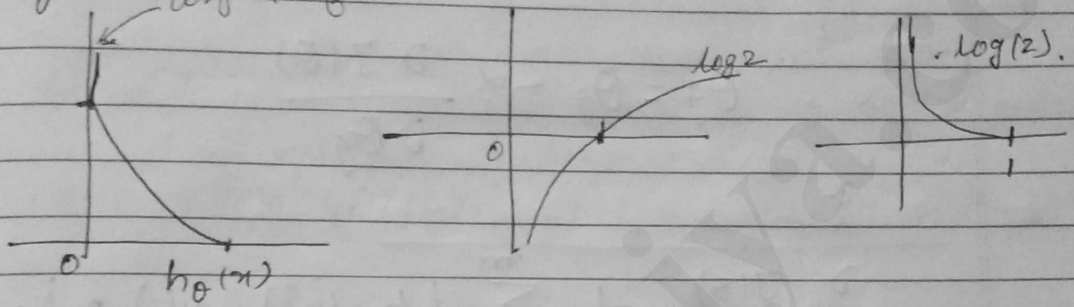
Page No.

Date ___/___/_____ Saathi

In case of logistic function we take cost function as :-

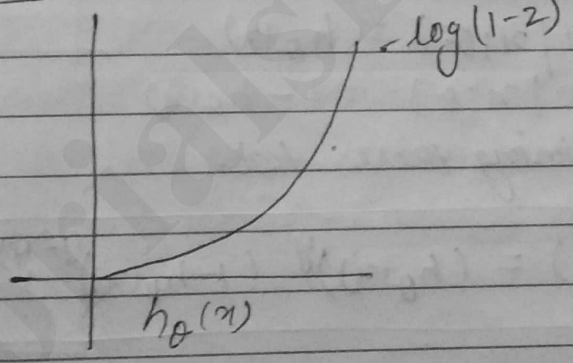
$$cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

If $y=1$



as value of z lies b/w 0-1 so we consider only the range b/w 0-1.

For $y=0$



$$cost(h_\theta(x), y) = -y \log(h_\theta(x)) - (1-y) \log(1-h_\theta(x))$$

Combined function for $y=0$ and $y=1$

when $y=0$

$$cost = -\log(1-h_\theta(x))$$

and

$y=1$

$$cost = -\log(h_\theta(x)).$$

Page No.

Saath

Date ___/___/_____

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} cost\,(h_\theta(x^i), y^i)$$

$$J(\theta) = \frac{-1}{m} \sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i) \log(1-h_\theta(x^i))$$

To find θ's

{

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$$

}

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)\, x^i_j$$

Maximum likelihood Approximation:-

$$P(y=1 \mid x;\theta) = h_\theta(x)$$
$$P(y=0 \mid x;\theta) = 1 - h_\theta(x)$$

On combining these two.

$$\boxed{P(y^i \mid x^i;\theta) = (h_\theta(x^i))^{y^i} (1-h_\theta(x^i))^{(1-y^i)}}$$

Likelihood

↳ $L = P(y \mid x;\theta)$

$$= \prod_{m} P(y^i \mid x^i;\theta)$$

$$L = \prod_{i=1}^{m} (h_\theta(x^i))^{y^{(i)}} (1-h_\theta(x^i))^{(1-y^i)}$$

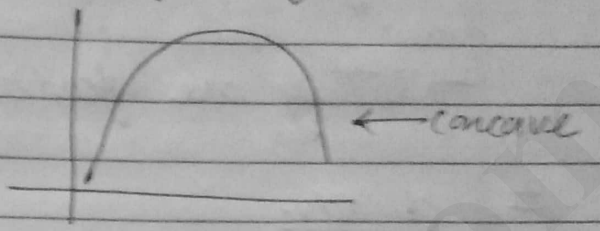
For mathematical simplicity we convert it in log function.

Date ___/___/_____

Saathi

log likelihood : $L.L(\theta) = \log L(\theta)$

$$L.L(\theta) = \sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^i)) + (1-y^i) \log(1-h_\theta(x^i))$$



But we want a convex graph because we want to minimize the function so add negative.

so.

$$L.L(\theta) = -\sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i) \log(1-h_\theta(x^i))$$

It is gradient ascent.

* Maximum likelihood estimation is a method of estimating the parameters of a statistical model given observations. The method obtains the parameter estimates by finding parameter values that maximumizes the likelihood function.

In practice it is actully convenient to work with natural logarithm of the likelihood function and we call it log likelihood.

26/02/19

Optimization Algo :-

→ Gradient Descent
→ Conjugate Gradient Descent
→ BFGS
→ L-BFGS

Saath

Date __/__/____

Adv. of other optimization algo. over gradient descent learning rate

→ you do not need to choose the value of α.

→ They are faster than gradient descent.

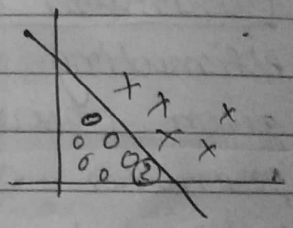
disadv.

They are more complex as compared to gradient descent.

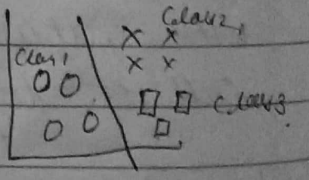
Multiclass classification

One Vs. All approach
or
One Vs. Rest.

If we have binary class so we can do binary classification



• But if more than two categories then we can't follow the above approach.

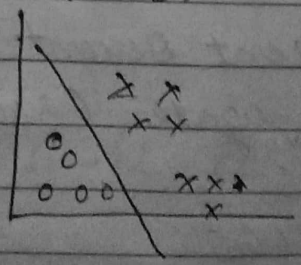


One Vs all or Rest

For checking class 1.

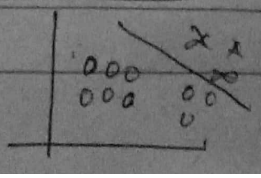
we just find the decision boundary for class 1. That is whether the particular can connect for class 1 or not.

Highest probability then we can say the training data belong to that class.



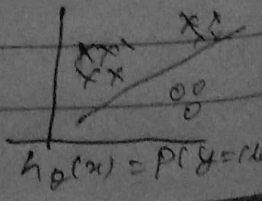
$h_\theta^1(x) = P(y = class\,1 | x; \theta)$

For class 2



$h_\theta^2(x) = P(y = class\,2 | x; \theta)$

For class 3



$h_\theta(x) = P(y = class\,3 | x; \theta)$

Date ___/___/_____

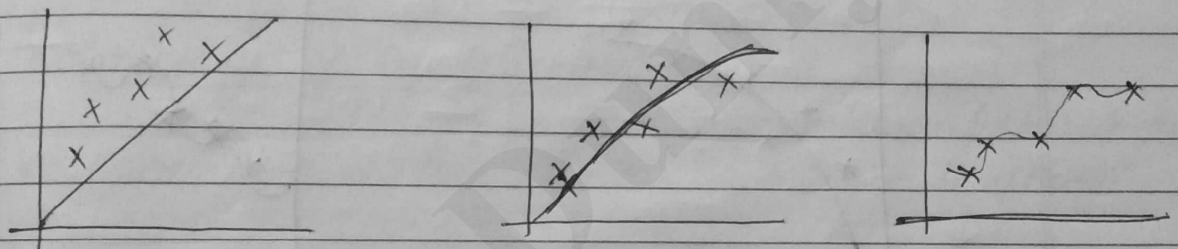
Saathi

Find. $\max_i h_\theta^{(i)}(x)$

Q. Suppose you have a multiclass classification problem with k classes. So. $y \in \{1 \dots k\}$ Using one Vs all method how many different logistic regression classifier will you endup training.

i) k-1
ii) k ✓
iii) k+1
iv) log(k)

Regularization



→ underfitting
→ (Model Oversimply due to my assumption (I assume that there is a linear relation)
→ Strong assumption
→ degree 1.

→ Normal
- degree of 2.

→ Overfitting
→ (degree of 1.2,...)
→ This model with high varience.
(due to lasts of feature than no. of example 120 classes.
→ High degree polynomial

Q. Consider a medical diagonosis of classifying tumors as meligrent or benige. If a hypothesis $h_\theta(x)$ has overfit the training set. It means which one of the following.

a) It makes accurate predication for example in training set and generalises when to make accurate prediction on new unseen example.

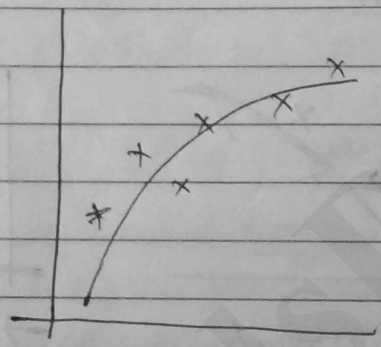
Saathi

Date __/__/____

(b) It doesn't make accurate for example in the training set, but it generalises well on new examples.

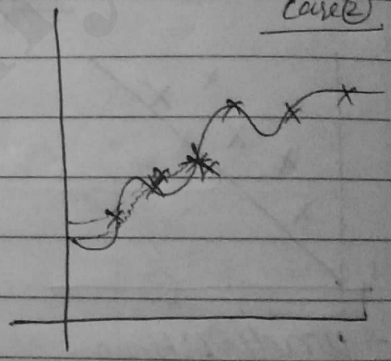
(c) It makes accurate prediction for examples in traing set but not generalises well on new eg.

(d) It doesn't make accurate predications for examp in traing set. and doesn't generalise to make prediction on new examples



$\theta_0 + \theta_1 x_1 + \theta_2 x_2$

overfitting.
case (c)



$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x^3 + \theta_4 x^4$

with Regularization we are trying to decrease the flexibility. So that reduce the varience and increase bias.

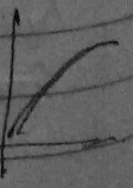
Best if (c) Take all attributes but reduce the coefficient of each attribute. which leads to normal form.

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

$$\downarrow 0 \quad \downarrow 0$$

$$= \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

↳ Quadratic



Date ___/___/___

Saathi

In case of overfitting:- Case (2)

$$\min \tilde{J}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + 1000 x_3^3 + 1000 x_4^2$$

If it have higher coefficient, so it also need to reduce at this same extent.

By reducing the coefficient the impact of that term also reduces in overall cost. which is possible only when we shrink the coefficient.

aim :- Complexity → simplicity.

28/2/19

Objectives of Regularization:-

To minimize / shrink the coefficients and thus, we are moving to simplicity.

Regularized Linear Regression

Regularized cost function:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right]$$

Regularization term

Regularization co-efficient

because θ_0 is not any feature thus, we begin it with 1.

Q. In regularized linear regression, we choose θ to minimize $J(\theta)$. What if λ is set to extremely large value, say $\lambda = 10^{10}$?

i. Algo works fine ii) Algo fails to eliminate over-fitting.

Saathi

Date __/__/____

(iii.) Algo results an underfitting.
(iv) Gradient descent will fail to converge.

iii) because to make the term we need to take value 0.0000000---- thus, we'll be left with only θ_0, thus it'll only make a line and all points will not be covered. Thus, underfitting.

Regularized Gradient descent:- will also change as cost func^n also changed.

Repeat {

will not change → $\theta_0 = \theta_0 - \frac{\alpha}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)$

$\theta_j = \theta_j - \frac{\alpha}{m} \left(\sum_{i=1}^{m} (h_\theta(x^i) - y^i) x_j^i + \frac{\lambda}{m} \theta_j \right)$

}

Taking θ_j common :-

$$\theta_j = \theta_j \left(1 - \frac{\lambda\alpha}{m}\right) - \frac{\alpha}{m} \sum_{i=1}^{m} \left[(h_\theta(x^i) - y^i) x_j^i \right]$$

this term < 1

thus θ is shrinking.

Normal Equation Method

$$\theta = (X^T X)^{-1} X^T y$$

In Regularization,

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{(n+1)\times(n+1)} \right)^{-1} X^T y$$

↓ no. of features.

Page No.

Date __/__/____

Saathi

The above takes care of non-invertibility as the matrix formed is always invertible.

Regularized Logistic Regression

$$J(\theta) = \frac{-1}{m}\left[\sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))\right] + \frac{1}{2m}\sum_{i=1}^{m}\theta_j^2$$

Gradient descent

repeat {

$$\theta_0 = \theta_0 - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_0^i$$

$$\theta_j = \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_j^i + \frac{\lambda}{m}\theta_j\right]$$

}

14/03/19

Confusion Matrix

A confusion matrix is a table often use to describe the performance of a classification model on a set of test data for which true values are known.

In this table the predictions are classified into four categories.

1. True Positive - actually output is +ve and predicted also positive.
2. True negative - actually output is false and predicted also false
3. False positive - actual out- negative & predictive +ve
4. False -ve - actual out +ve and

TutorialsDuniya.com

Download FREE Computer Science Notes, Programs, Projects, Books PDF for any university student of BCA, MCA, B.Sc, B.Tech CSE, M.Sc, M.Tech at https://www.tutorialsduniya.com

- Algorithms Notes
- Artificial Intelligence
- Android Programming
- C & C++ Programming
- Combinatorial Optimization
- Computer Graphics
- Computer Networks
- Computer System Architecture
- DBMS & SQL Notes
- Data Analysis & Visualization
- Data Mining
- Data Science
- Data Structures
- Deep Learning
- Digital Image Processing
- Discrete Mathematics
- Information Security
- Internet Technologies
- Java Programming
- JavaScript & jQuery
- Machine Learning
- Microprocessor
- Operating System
- Operational Research
- PHP Notes
- Python Programming
- R Programming
- Software Engineering
- System Programming
- Theory of Computation
- Unix Network Programming
- Web Design & Development

Please Share these Notes with your Friends as well

4 Cabin

Date __/__/__

actual \ Predicted	T	F
T	TP	FN
F	FP	TN

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}}$$

$$\text{Mis classification error} = \frac{FP + FN}{\text{Total}}$$

Precision → when predicted yes, how often it is true / correct

$$\frac{TP}{\text{Predicted Yes}}$$

Q. Consider a training of a classifier on a total of 500 rows of training data. To predicte a class variable as yes or no. The class variable could predicte correct yes for 230 no. of rows, incorrect no for 120 no. of rows, incorrect yes for 100 no. of rows and incorrect no. for 50 no. of rows. Make a confusion matrix. Calculate accuracy and mis-classification error.

Soln:

actual \ Predicted	T	F
T	230	50
F	100	120

- Accuracy $= \frac{230+120}{500} = \frac{350}{500}$

$= \frac{7}{10} = 0.7$

- Mis $= \frac{150}{500} = \frac{3}{10} = 0.3$

Date __/__/____

Saathi

$$g(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$$

$$g'(z) = \frac{(1+e^z)e^z - e^z(e^z)}{(1+e^z)^2}$$

$$= \frac{1}{(1+e^z)} \times e^z$$

$$\text{or } g'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} \times \frac{e^{-z}}{1+e^{-z}}$$

$$= \left(\frac{1}{1+e^{-z}}\right)\left(1 - \frac{1}{1+e^{-z}}\right)$$

$$= g(z)\,(1-g(z))$$

so,

$$\boxed{g'(z) = g(z)\,(1-g(z))}$$

this is the property of logistic function that its derivative comes in its own term.

H.M Bipolar sigmoidal function and its derivation

Q.1 Differentiate between classification and clustering technique and their applicability with example of both.

- Classification

Neural Networks

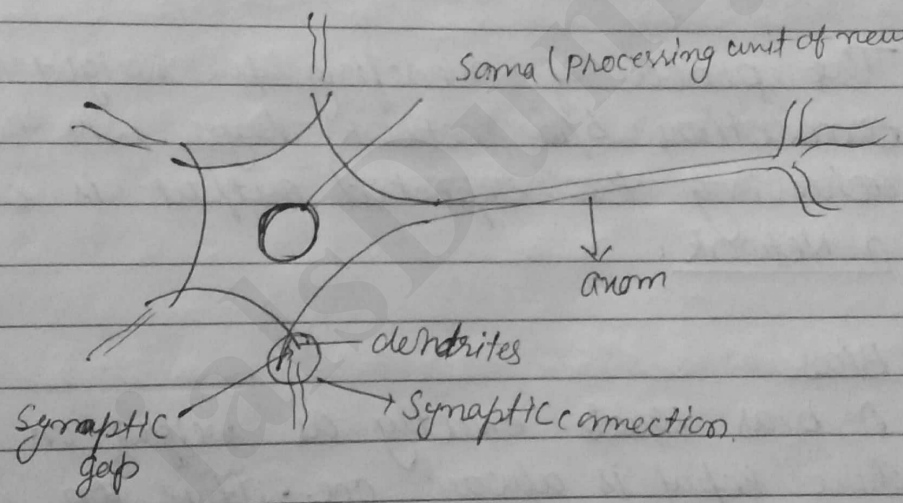
Date 29/03/19

Saathi

Q Diff biological Neurah & Artificial Neural Network.

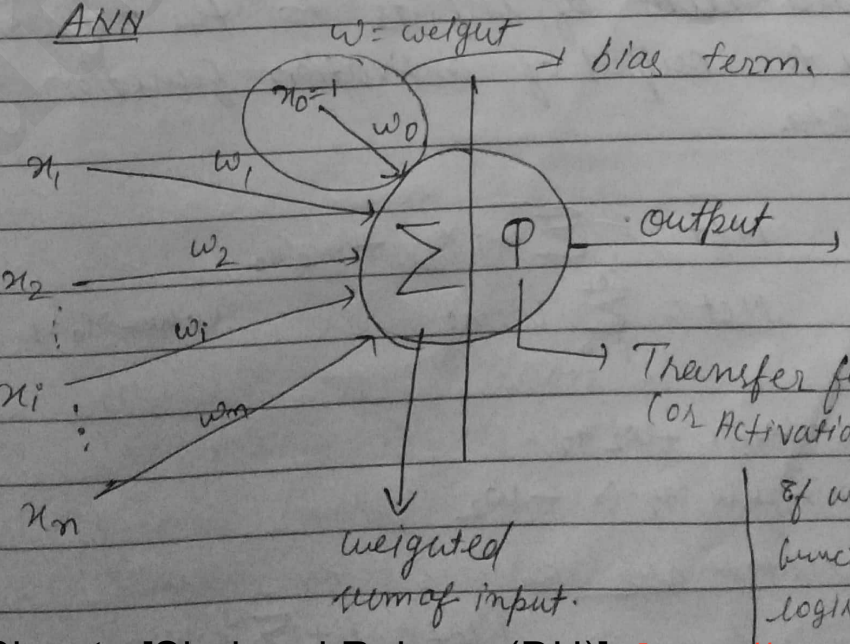
- Artificial Neural Networks are information built from interconnected elementry processing unit known as Neuron. It is inspired by the way biological nervous systems such as human brain process information. ANN provides a general, practical method for learning real value, discrete value functions from examples. ANN consists of many nodes each node is analegous to neuron in the human brain. Each node has no. of inputs, associated node function and an output.

fig :-



Biological Neuron

fig. ANN



If we use φ as sigmoidal function the work as logistic use for classification

Saathi

Date ___/___/_____

- **Weight :**
For a Neural Network, Neurons are connected to each other, using connection links. Each connection link is associated with a weight. Weight is the info used by NN to solve a problem.

- **Net :**
is the summation of products of weights and input signals inccedenting on a neuron.

$$Net = \sum w_i x_i$$

- The process of modefing the weight in the connections b/w network layer with the objective of achieving the expected output is called Training a Network.

- **Bias**
A bias equals exactly as weight on a connection whose input is always one. It is an additional parameter use to adjust output along with the weighted sum of inputs to the neuron. It allows shifting the output of activation function to better fit the data.

$$Net = \sum w_i x_i + w_0 x_0$$

$$Net = \sum_{i=0}^{n} w_i x_i \qquad \text{when } x_0 = 1$$

$$w_0 x_0 + w_1 x_1 + w_2 x_2$$

$$w_1 x_1 + w_2 x_2 > -w_2$$

Net is also as

$$Net = b + \sum_{i=1}^{n} x_i w_i$$

Page No.

Date __/__/____

Saathi

Activation fun^n.

Activation funt^n converts input signal of a note in ANN to the output signal. It decides whether a neuron should be activated or not. the purpose of activation function is to bring non-linearity into output of neurons. The activation function does the non-linear transformation of the input making it capable to learn and perform more complex task

$$\phi(z) = z$$

◉ Different types of activation function :-

- **Linear Activation function.**
  $f(x) = ax$
  range = $-\infty$ to ∞
  use - suitable to use at o/p layer.

- **Sigmoidal function**
  $f(x) = \frac{1}{1+e^{-x}}$
  range 0-1
  use : o/p layer of binary classification problem.

- tan h - mathematically shifted version of sigmoidal function.

$$f(x) = \frac{2}{1+e^{-2x}} - 1$$

Range :- -1 to 1
→ used in hidden layers of a neural network as its value lies b/w -1 and 1 Hence helps in centering of data by bringing mean close to 0 and simplifying learning next layer.

Saathi

Date

* $sgn(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{otherwise.} \end{cases}$

Diff

ANN	Biological Network N/w
1.) Processing speed y 10^{-1}s	→ Processing speed 10^{-3}s
→ Parallelism - comparatively less ||.	→ It is massively || architecture
→ size & complexity donot involve this much computational.	→ estimated to contain 10^{11} neurons each connected to 10^4 other neurons
→ New info destroys the old info.	→ New info doesn't destroy the old info.
→ individual unit in ANN o/P a single const. value.	→ outputs a complex time-series of spikes

Q. Explain system ALVINN - prototype of ANN

Date 3/04/19

Saathi

Characteristics :- of Back-Propagation :-

1) Instances are represented by many attribute-value pairs.

2) Target function output may be real value, descrete valued or a vector of several real or discrete values.

3. The training examples may contain errors

4. Long training times are acceptable.

5. Fast evaluation of learned target function, may be required.

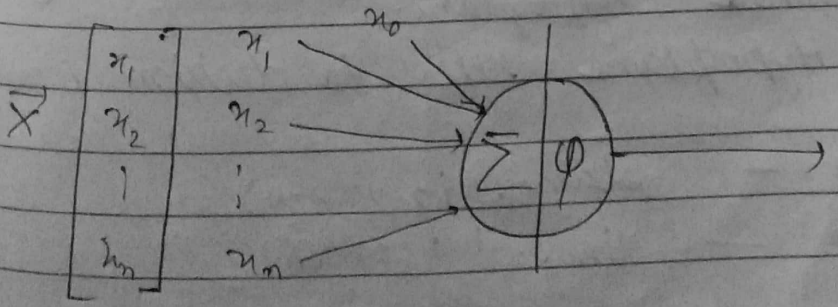
6. The ability of humans to understand the learned target function is not important.

Perceptron

A unit of neural network that takes a vector of real value input, calculate the weighted sum or linear combination of inputs, then output a 1 if result is greater than some threshold and -1 otherwise.

$$O(x_1, x_2, \ldots x_n) = \begin{cases} 1 & \text{if } w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n > 0 \\ -1 & \text{otherwise.} \end{cases}$$

also called sgn.
(output is 1 & -1)



Page No.

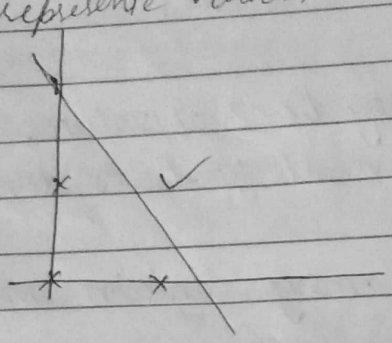
Saathi

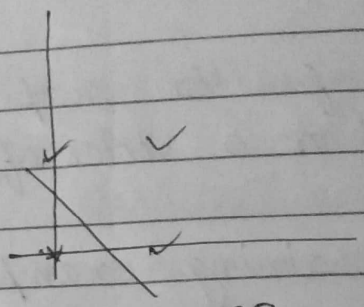
Date __/__/____

imp Q. Architecture of Perceptron?

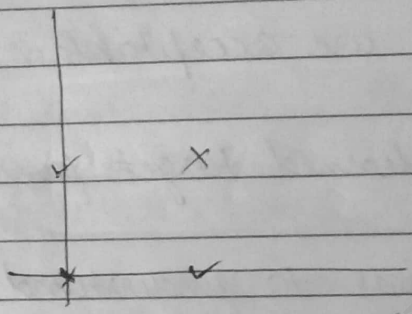
Problem associated with single layered Perceptron:
represents various boolean functions.

For AND





OR



XOR (Can't make separate these with help of only one decision line).

Perceptron won't be able

Define the
Q. Concept of linear separability / proble with perceptron — (XOR)

Linear Separability:
Is the concept to check if (output) points of two diff. classes in n dimensions can be separated by an (n-1) dimensional hyperplane.

Here Hyperplane refers to decision boundary

3-D — plane. (3-1 = 2-D)

2-D — line. (2-1 = 1D).

Date ___/___/_____

Saathi

Various Rule to learn the weight of perceptron:

1) Perceptron training rule.
2) Delta Rule.
3) Linear Programming.

1. Perceptron training Rule :

i) Begin with random weights and iteratively apply perceptron to each training example.

ii) Modify perceptron weights whenever it mis-classify an example.

iii) This process is repeated as many times as required untill the perceptron classifies all training example correctly.

Perceptron training Rule : $w_i \leftarrow w_i + \Delta w_i$

$\Delta w_i = \eta (t - O) x_i$

(x_i → input; O → current output; t → target out; η → learning rate)

Perceptron training rule is proven to converge within a finite number of applications of this rule to weight vector that correctly classifies all training examples provided the examples are linearly separable.

2. Delta Rule :

$$w_i \leftarrow w_i + \Delta w_i$$

$$\Delta w = -\eta \nabla E(\vec{w})$$

Error function : $E = \frac{1}{2} \sum_{d \in D} (t_d - O_d)^2$

D :- set of all examples

d :- one example from D.

Date __/__/____

$$E = \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2$$

$$= \frac{1}{2} \sum_{d \in D} 2 (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d)$$

$o_d = \vec{w} \cdot x_d$ (linear comb.)

$$= \sum_{d \in D} (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - \vec{w} \cdot x_d)$$

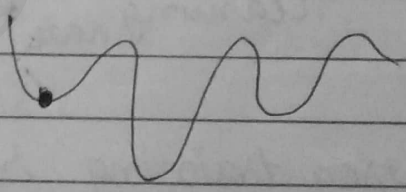
$$= -\sum_{d \in D} (t_d - o_d) x_{id}$$

On substituting

$$\boxed{\Delta w = \eta \sum_{d \in D} (t_d - o_d) x_{id}}$$: delta Rule.

problem

- This process might be slow. (overcome by stochastic proble
- stuck at local minima.



Q. Diff b/w Δ rule and perceptron rule.

Q. Diff b/w. Standard gradient decent and stochastic gradient descent.

Standard Gradient descent	Stochastic Gradient descent
1 Error is summed over all the example before updating the weights	: Weights are updated upon examining each training example

Page No.

Date __/__/____

Saathi

1. Summing over all or multiple example in Standard gradient descent requires more computation per weight update step.	2. Less computation is required per weight update step.
3. It uses the true gradient descent. Therefore it use larger step size per weight update	3: Smaller step size per weight update step.
4. It uses gradient of error $\nabla E(\vec{w})$ to guide search and hence more chances to fall in local minima.	4. It uses gradient of error $(\nabla E_d(\vec{w}))$ to guide the search. The parameters as estimated for every observation as oppose to whole samples. in, It reduces a lot of randomness. the path of this vendors over more places and thus avoid stucking into local minima.
5. $\Delta w_i = \eta \sum_{d \in D} (t_d - o_d) x_{id}$	5. $\Delta w_i = \eta (t - o) x_i$

Q. Diff. b/w Δ rule & perceptron rule :-

Perceptron rule	Δ rule
1. This rule updates the weight based on output from a thresholded perceptron	1. This rule updates the weight based on unthresholded perceptron.
2. output: $o = sgn(\vec{w} \cdot \vec{x})$	2. output $o = \vec{w} \cdot \vec{x}$

Date __/__/____

Saathi

3. Convergence property perceptron training rule converges after finite number of iterations to a hypothesis that perfectly classify the training data provided training examples are linearly separable.

3. Δ rule converges only asymptotically towards minimum error hypothesis possibly requiring unbounded time but converges regardless of whether training data is linearly separable or not

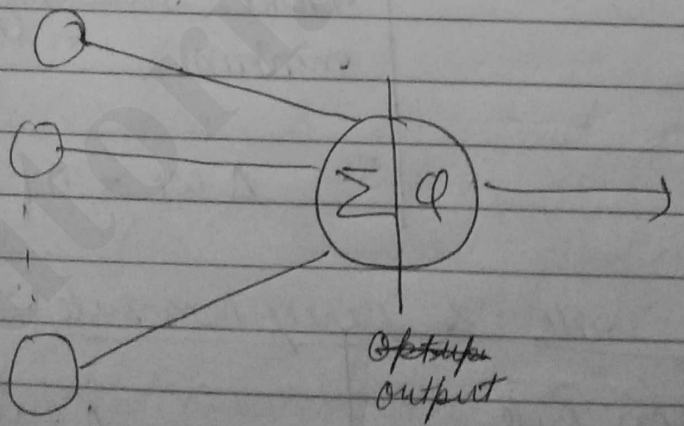
4.

(3) Linear Programming :-

→ It can't be extended for multi layer examples.

→ Use only for linearly separable type example.

- Single layer



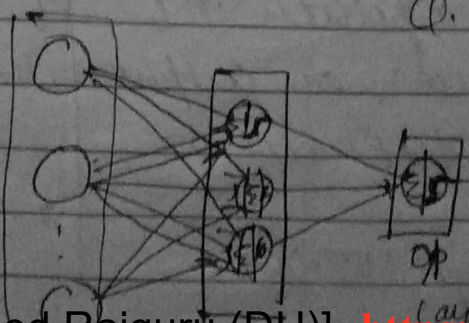
Input layer.

- Multilayer

— No reverse link (so called feed forward network)

bt error are back propagated

Q. Draw a diagram for multilayer perceptron



Saathi

We can solve non-linearly separable problems with the help of multi-layer perceptron.

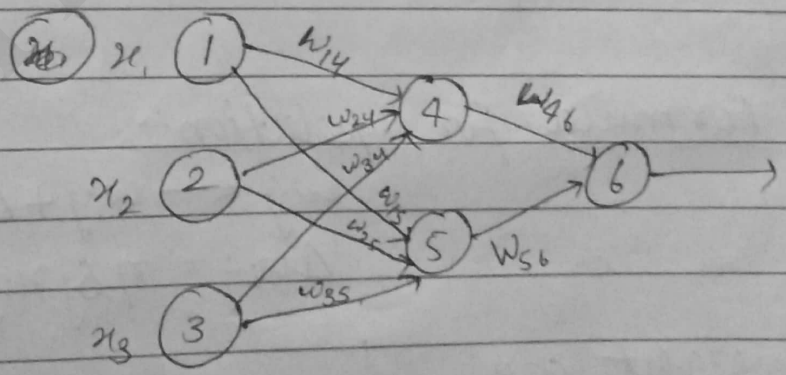
A unit in multilayer architecture should have:-

1. Output is non-linear function of inputs.
2. Output is differentiable function of input.

9/04/19

Q. Consider a multi-layer feed forward network as given in figure. Let the learning rate be 0.9. The initial weights and bias values are given in table along with first training tuple $X = (1, 0, 1)$ with class label 1.

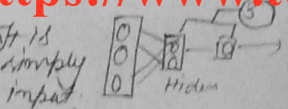
1.) Calculate net input and output of each neuron.
2.) Calculate error at each node.
3.) Calculate updated weight and bias.



x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	-0.3	0.4	0.1	-0.5	0.2	-0.3	-0.2	-0.4	0.2	0.1

Bias - x_0 (w_0) value $x_0 = 1$ to θ_4 No. of neuron = 3.

Bias - θ_4



Saathi

Date __/__/____

Solution:

(1)

Unit	Net Input	O/P (If sigmoid)
4	$0.2 + 0 - 0.5 + 1(-0.4)$ (bias [x_0 & w_0] here θ_4) $= -0.7$	$\frac{1}{1+e^{0.7}} = 0.332$
5	$-0.3 + 0 + 0.2 + 0.2 = 0.1$	$\frac{1}{1+e^{-0.1}} = 0.525$
6	$-0.3 \times 0.332 + 0.525 \times -0.2 + 0.1$ $= -0.105$	$\frac{1}{1+e^{0.105}} = 0.474$

(2)

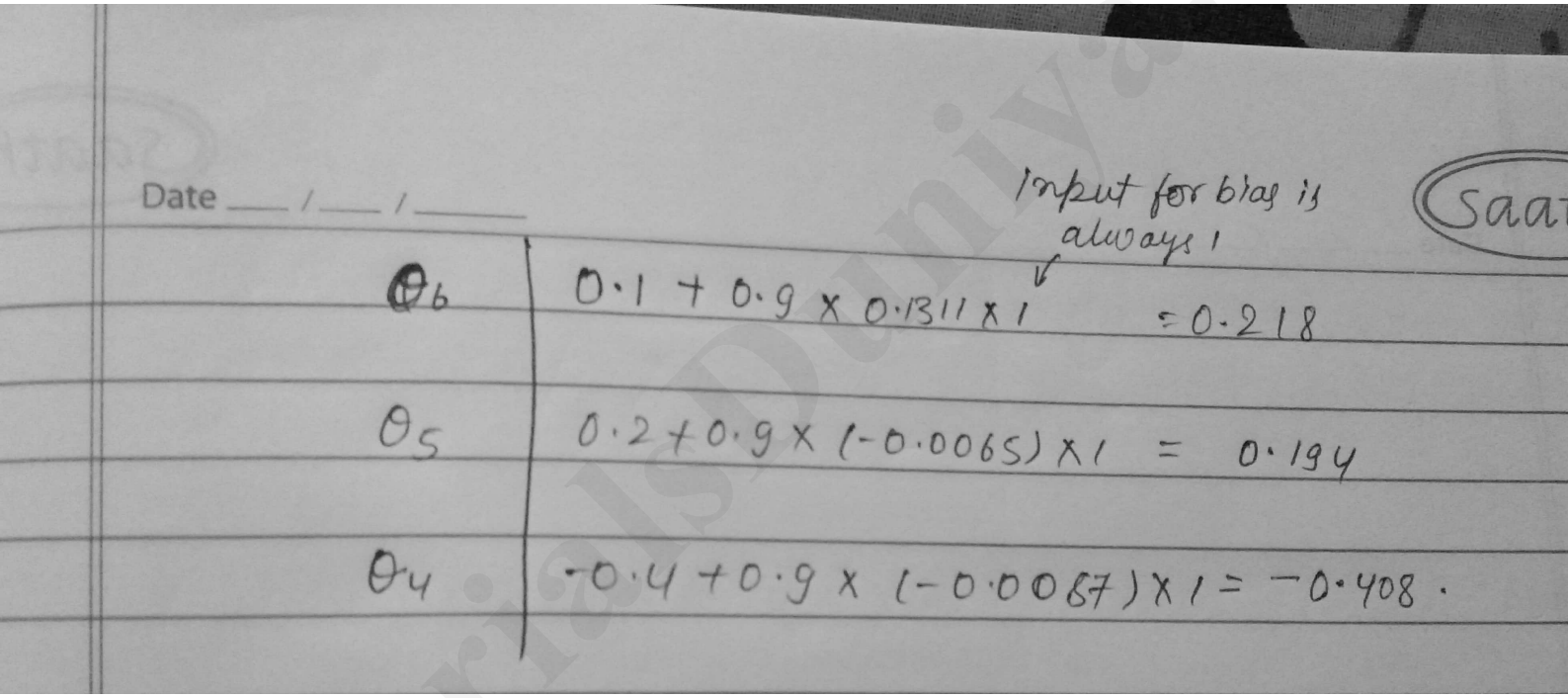
Unit	Error	Formula
		• For output layer
6	$0.474(1-0.474)(1-0.474)$ $= 0.1311$	$\delta_k = O_k(1-O_k)(t_k - O_k)$
		• For hidden layer
5.	$0.525(1-0.525) * (-0.2) \times 0.1311$ $= -0.0065$	$\delta_h = O_h(1-O_h)\sum_{k \in \text{o/p}} w_{hk}\delta_k$ (Secondary)
4	$0.332(1-0.332) \times (-0.3) \times 0.1311$ $= -0.0087$	

(3) Formula for updation:-

$$w_{ij} \leftarrow w_{ij} + \Delta w_{ij}$$

$$\Delta w_{ij} = \eta \delta_j x_{ij}$$

weight or Bias	New value
w_{46}	$-0.3 + 0.9 \times 0.1311 \times 0.332$ $= -0.261$
w_{56}	$-0.2 + 0.9 \times 0.1311 \times 0.525 = -0.131$
w_{14}	$0.2 + 0.9 \times (-0.0087) \times 1 = 0.192$
w_{15}	$-0.3 + 0.9 \times -0.0065 \times 1 = -0.306$
w_{24}	$0.4 + 0.9 \times -0.0087 \times 0 = 0.4$
w_{25}	$0.1 + 0.9 \times (-0.0065) \times 0 = 0.1$
w_{34}	$-0.5 + 0.9 \times (-0.0087) \times 1 = -0.508$
w_{35}	$0.2 + 0.9 \times (-0.0065) \times 1 = 0.194$



Date ___ / ___ / _____

Input for bias is always 1

θ_6	$0.1 + 0.9 \times 0.1311 \times 1 = 0.218$
θ_5	$0.2 + 0.9 \times (-0.0065) \times 1 = 0.194$
θ_4	$-0.4 + 0.9 \times (-0.0087) \times 1 = -0.408$