

TUTORIALSDUNIYA.COM

Machine Learning Notes

Contributor: Shweta
[Shaheed Rajguru DU]

Computer Science Notes

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

Please Share these Notes with your Friends as well

facebook

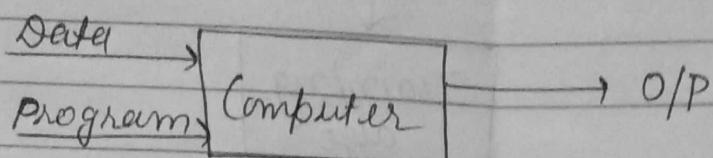


MACHINE LEARNING

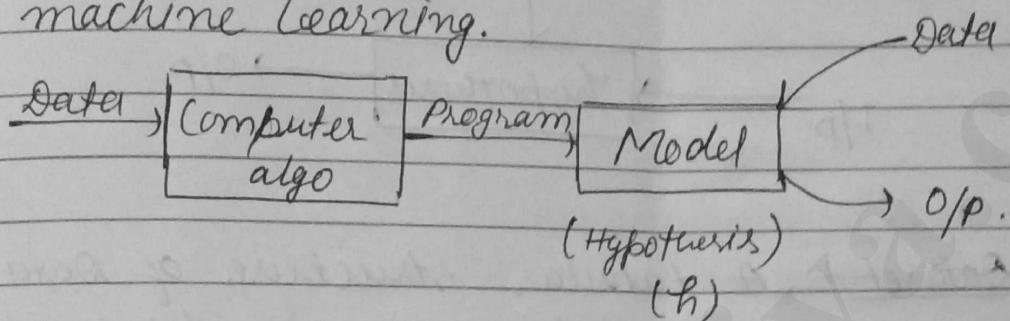
Date 2, 01, 19

Saathi

In traditional computer.



In machine learning.



e.g. House Price Prediction.

	Area (sq.ft)	Price
i)	100	—
ii)	102	—

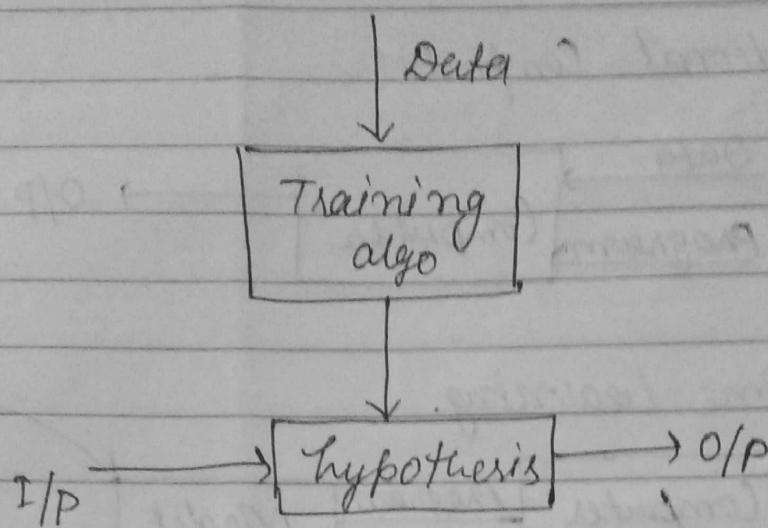
↓
Pays this data to algo, then
machine finds a unique mapping.

Definition :

Computer program is set to learn from experience (E) with respect to some class of Task (T) and Performance Measure (P), if its performance at Task T as measured by P improves with experience (E).

- eg:
- spam filtering in gmail.
 - Face recognition
 - Google assistance
 - Weather forecasting,

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.

The diagram shows a table structure representing a dataset. The columns are labeled "Area", "Color", "# of rooms", and "Price". The "Price" column is circled and labeled "O/P label / response variable / dependent variable". A bracket on the left labeled "Examples/instances" spans all four columns. A bracket below the first three columns is labeled "features / Independent variable".

Area	Color	# of rooms	Price

Training Data 80% of total Data is use to train 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.

Types of M/L :-

- i) Supervise learning.
- ii) Unsupervised learning - clustering
- iii) Reinforcement learning.

Supervise 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 response's value. It is also known as learning with examples. Algorithm here tries to find out the relationship between input values and output values.

Supervise Learning categorises into:

- i) Regression problem
- ii) Classification problem.

Data is represented as -

• Supervise

$$D = \{(x^i, y^i)\}_{1}^N$$

↳ corresponding Response.

N :- No. of Rows.

• Unsupervised

$$D = \{x^i\}_{1}^N$$

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 somehow similar or related to each other based on feature like

Date _____

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

- Clustering Problem

Representation of data in Supervise learning.

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ \vdots \\ x^{(m)} \end{bmatrix}_{m \times 1} = \begin{bmatrix} x_1^1 & x_2^1 & x_3^1 & \cdots & x_n^1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^j & x_2^j & x_3^j & \cdots & x_n^j \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^m & x_2^m & x_3^m & \cdots & x_n^m \end{bmatrix}_{m \times n}$$

(input matrix)

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

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

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

In regression:

$$f(x) : X \rightarrow Y$$

$$X \in \mathbb{R}^m$$

$$Y \in \mathbb{R}^1 \quad [\text{bcz } Y \text{ is single dimensional}]$$

Classification:

$$f(x) : X \rightarrow Y$$

$$X \in \mathbb{R}^m$$

$$Y \in \{1, \dots, k\}$$

Date 8, 01, 19

Application of ML :-

1) Learning Associations

↳ supermarket for Basket Analysis

- 2) Face Recognition
- 3) Fraudulent transactions
- 4) High risk / low risk
- 5) pattern recognition

↳ Handwriting recognition

Machine Learning work flow:-

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

$$P(B \cap A) = P(B|A) \cdot P(A)$$

$$P(A \cap B) = P(A|B) \cdot P(B)$$

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

[]

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

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

Date _____ / _____ / _____

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

(H)

$$h_i \quad P(h_i | D) = \frac{P(D|h_i) \cdot P(h_i)}{P(D)}$$

likelihood

prior probability
of hypo

posterior
probability

evidence / probability of
current data.

h_{MAP} = argmax posterior prob

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

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h) \cdot P(h)}{P(D)}.$$

as D become same
for every h_1, h_2, \dots
to drop $P(D)$

~~for hypothesis~~ $h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(D|h) \cdot P(h)$

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

Lets. $P(h_1) = P(h_2) = \dots$
to, 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 _____

E1.2.3

- Q. Consider a medical diagnostic problem. There are two alternative hypothesis.
1. The patient has a particular form of cancer.
 2. Patient doesn't have cancer.

The lab test which is used to diagnose 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 diagnose the patient has cancer or not.

Ans:-

h_1 = Person has cancer.

h_2 = Person doesn't have cancer.

Find :-

$P(h_1 | D)$?

D: Result of the lab

$P(h_2 | D)$?

test is +ve.

$$P(\text{cancer}) = 0.008$$

$$P(7\text{cancer}) = 0.992$$

$$P(+|\text{cancer}) = 0.98$$

$$P(-|\text{cancer}) = 0.02$$

$$P(-|\text{not cancer}) = 0.97$$

$$P(+|7\text{cancer}) = 0.03$$

Final

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

$$P(7\text{cancer} | +)$$

$$P(+|\text{cancer}) \times P(\text{cancer})$$

$$h_1 =$$

$$= 0.98 \times 0.008$$

$$= 0.0078$$

Saath

Date _____

$$h_2 = (+ | \text{cancer}) \times P(\text{cancer}) \\ = 0.03 \times 0.992 \\ = 0.0298$$

Here, $h_2 > h_1$ So, $h_{MAP} = \text{cancer}$

Now, to change it into 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 filtration

document classification etc

It is based on Bayes theorem and is a probabilistic classifier.

Date _____

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

V_{map}

$$V_{map} = \operatorname{argmax}_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n)$$

$$= \operatorname{argmax}_{v_j \in V} \frac{P(a_1, a_2, \dots, a_n | v_j) \cdot P(v_j)}{P(a_1, a_2, \dots, a_n)}$$

$$= \operatorname{argmax}_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j) \cdot P(v_j)$$

$$V_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \times \prod_{i=1}^n P(a_i | v_j)$$

$$\left[P(a_1, a_2, \dots, a_n) = \prod_{i=1}^n P(a_i | v_j) \right]$$

9/01/19

Lab

esc + m → code cell to markdown // use to heading

esc + y → code cell from markdown.

shift + enter → 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.

Page No.

Saath

Date _____

end = '' → to change new line to something
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 ("{} loves {}".format (a, b))

Output → Indians loves mangoes

Date _____

Saathi

Data type

primitive - int
advance - list, tuple

Operations

+,-, ÷, floor, division, **(exponent)

- print ("%.0.4f", (a/b))
→ round off till four position.
- Multiple variable assignment
 $a, b, c = 10, 20, 30;$
- Comment
 - single #
 - multiline " " " " " "
- Markdown executed while comments not.
- whitespaces
- for i in range (1, 10)
print (i) → starting index.
1
2
3
4
5
6
7
8
9
→ Jumping step
- for i in range (1, 10, 2)
1
3
5
7
9
→ Reversed.
- for i in range (10, 1, -1).

Saathi

Date function

def myfunc (name, 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

print (a[0:2])

print (a[0:]). // whole

print (a[0:-2]). // MAN.

print (a[-1:]) // O

a = "MANGO"

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]

0010
0100
0110

Saathi

Date _____

print(type(s)) // 1st.

print(type(l[0])) // 2nd.

- a. `splittlines()` // split at new line.
- `upper()` — to convert in uppercase.
- strip fun " to remove leading & trailing spaces.
- isdigit for check only digit.
- isalnum for alphabet + number have.
- `find()` fun // return index number for multiple. from left and reverse
- `rfind()`
+ running from right and return its index number.
- `count()` + occurrence. count.
- " ".join(l);
↑ output : my-Python.
Join choose.
- a. `capitalize()` → to capitalize 1st letter.
- a. `title()` → My Python.
- capitalize every words 1st lett
- Logical operator

2^4

0010

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

facebook

WhatsApp 

twitter 

Telegram 

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 Bayes classifier on the above data to classify the following Novel instance

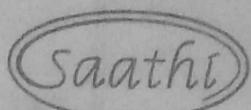
$X = \{ \text{outlook} = \text{sunny}, -\text{Temperature} = \text{cool}, \text{Humidity} = \text{high}, \text{wind} = \text{strong} \}$

Ans.

Naive Bayes Expression

$$V_{NB} = \arg \max_{V_j \in V} P(V_j) \times \prod_{i=1}^n P(a_i | V_j)$$

Date ___ / ___ / ___

Case 1 $V_1 = \text{Yes}$.

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

$$= \frac{9}{14} \times \frac{2}{8} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = \frac{1}{63 \times 3} = \frac{1}{189}$$

Case 2 $V_2 = \text{No}$

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

$$= \frac{5}{14} \times \frac{3}{8} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = \frac{36}{1750} \quad \frac{852}{125}$$

Here, $V_2 > V_1$

So,

$$V_{\text{NB}} = V_2.$$

Date _____ / _____ / _____

Saath

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 attribute buys computer has two distinct values yes/no. The tuple x is as follows.

$x = \{age = youth, income = medium, student = yes, credit rating = fair\}$

Let

$V_1 = Yes$.

$P(yes) \times P(age = youth | yes) \times P(income = medium | yes)$
 $\times P(student = yes | yes) \times P(credit rating = fair | yes)$

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

Date $V_2 = 1/V_0$

$$P(V_0) \times P$$

$$\begin{array}{r} 125 \\ - 75 \\ \hline 473 \\ - 31 \\ \hline 16 \end{array}$$

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

$$\text{a), } V_1 > V_2$$

$$V_{NB} = V_1$$

- Q. Consider the following six following training instances.

STATUS	FLOOR	DEPARTMENT	OFFICE SIZE	MEMBERS	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 classifier would classify Recycle Bin attribute for the following instance.

$X = \langle \text{status} = \text{student}, \text{floor} = 4, \text{department} = \text{CS}, \text{office size} = \text{small} \rangle$

③ Find the likelihood of recyclebin for the instance below

$X = \langle \text{status} = \text{faculty}, \text{floor} = 2, \text{department} = \text{CS}, \text{office size} = \text{large}, \text{members} = 97 \rangle$

(Saathi)

Date / /

- ① Predicting table
→ For status.

→ for Floor

STATUS	YES	NO	Floor	YES	NO
Faculty	2/4	0/2	3	2/4	0
Student	2/4	0/2	4	2/4	1/2
Staff	0/4	2/2	5	0/4	1/2

→ for Department

→ for Office size

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

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

2. Given instance

$X = \langle \text{status} = \text{student}, \text{floor} = 4, \text{department} = \text{CS}, \text{office size} = \text{small} \rangle$

→ For $V_1 = \text{yes}$

$$P(\text{yes}) \times P(\text{status} = \text{student} / \text{yes}) \times P(\text{floor} = 4 / \text{yes}) \times P(\text{department} = \text{CS} / \text{yes}) \\ \times P(\text{office size} = \text{small} / \text{yes})$$

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

→ For $V_2 = \text{No.}$

$$P(\text{No}) \times P(\text{status} = \text{student} / \text{No}) \times P(\text{floor} = 4 / \text{No}) \times P(\text{department} = \text{CS} / \text{No}) \\ \times P(\text{office size} = \text{small} / \text{No})$$

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

Saathi

Date _____

as. $V_2 > V_1$

Hence

$$V_{NB} = V_2$$

3. Given instance :-

$x = \langle \text{status} = \text{faculty}, \text{floor} = 2, \text{department} = \text{CS}, \text{officesize} = \text{large}, \text{member} = 97 \rangle$

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)
```

Date _____ / _____ / _____

Saath

List

- as an array
- heterogenous elements.

* List of square

$L4 = [i^2 \text{ for } i \text{ in range}(1, 6)]$
print(L4)

O/P → [1, 4, 9, 16, 25]

* List slicing

print(L4[0:4])

print(L4[-1])

O/P. [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 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 _____

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 dictionary.

To get all the keys.

`print(d.keys())`

all values

`print(d.values())`

`print(d["mango"])` # It return error if not exist
in the dict

`d.get("mango")` # It return None if it is not
present in dict.

`l1 = ["apple", "papaya", "grape"]`

`l2 = [100, 200, 300]`

`p = zip(l1, l2)`

Practical
4 Q.

Date _____

Saati

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.

e.g.

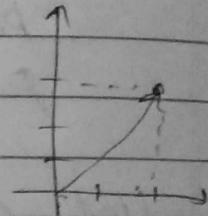
5 km/h speed → scalar

5 km/h in east → velocity.

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

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

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



- vectors are always ordered

means

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

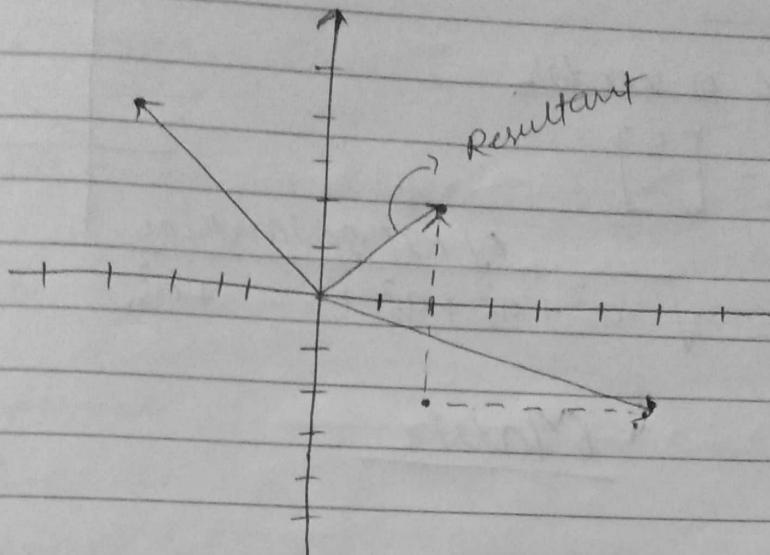
Addition of vectors :-

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

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

Date _____ / _____ / _____

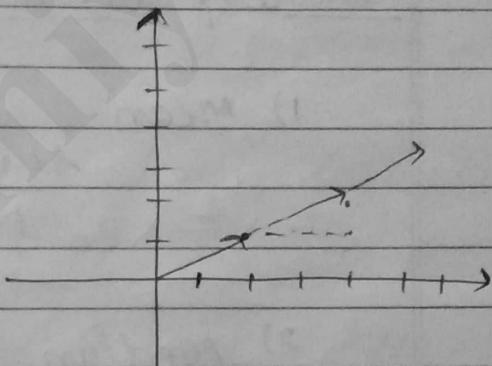
Saathi



Multiplication of vector

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

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



Linear combination of a vector

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

$$c_1 v_1 + c_2 v_2 + c_3 v_3 + \dots + c_n v_n \quad [\text{linear combi}]$$

Dot product

$$\vec{a} = \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vdots \\ \vec{a}_m \end{bmatrix} \quad \vec{b} = \begin{bmatrix} \vec{b}_1 \\ \vec{b}_2 \\ \vdots \\ \vec{b}_n \end{bmatrix}$$

$$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 + a_3 b_3 + \dots + a_n b_n$$

$$\text{Also, } \vec{a} \cdot \vec{b} = \vec{a}^T \cdot \vec{b} = [a_1 \ a_2 \ \dots \ a_m] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Date _____

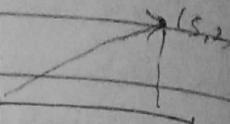
Saathi

Length of a vector.

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

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

by pythagorean theorem

MatrixStatMeasure 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

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

2) Standard Deviation

$$= \sqrt{\text{Variance}}$$

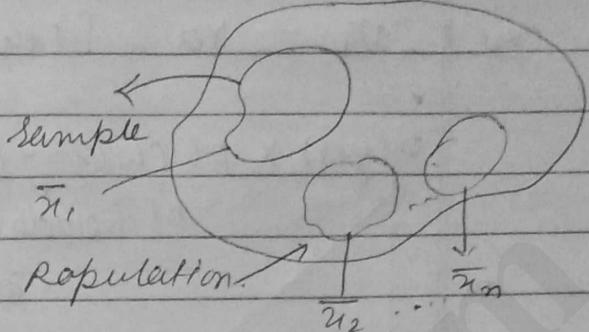
$$= \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$$

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

Date _____ / _____ / _____

For mean (u).

- 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 patien patients I have seen, below are their symptoms and diagnosis.

	chills	runny 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, runny nose = N, headache = Mild, fever = Y, flu = ?}

$$P(A|D) = \frac{P(D|A) \times P(A)}{P(D)}$$

Saath,

Date _____

(Case 1) $V_1 = \text{Yes}$ [Flu is yes]

$$P(\text{yes}) \times P(\text{chills} = \text{y} / \text{yes}) \times P(\text{runny nose} = \text{N} / \text{yes}) \times \\ P(\text{headache} = \text{mild} / \text{yes}) \times P(\text{fever} = \text{y} / \text{yes})$$

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

(Case 2) $V_2 = \text{No}$

$$P(\text{No}) \times P(\text{chills} = \text{y} / \text{No}) \times P(\text{runny nose} = \text{N} / \text{No}) \times \\ P(\text{headache} = \text{mild} / \text{No}) \times P(\text{fever} = \text{y} / \text{No})$$

$$= \frac{3}{8} \times \frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{108}$$

$$V_1 > V_2 \\ \text{so, } \boxed{V_{\text{NB}} = V_1}$$

Hence it classify that patient shows the symptom has flu.

$$P(V_1) = \frac{\frac{3}{125}}{\frac{3}{125} + \frac{1}{108}} = \frac{\frac{3}{125}}{\frac{125 \times 108}{108 \times 3 + 125}} = \frac{3}{324+125} \\ = \frac{324}{449}$$

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

UNIT -3

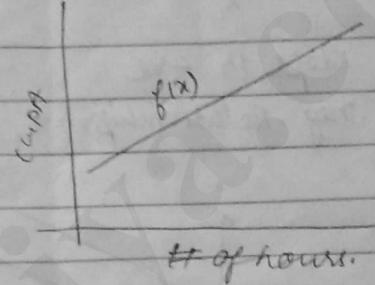
Date 22, 01, 19

Linear Regression

Saathi

- O/P variable is continuous then regression problem.
- O/P variable is discrete then classification problem.
- Under Regression problem, we plot the data.

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

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 / Predictor
variable
(may be one or more)

Simple

↳ multiple.

Page No. 1

↳ explanatory
variables

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

facebook

WhatsApp 

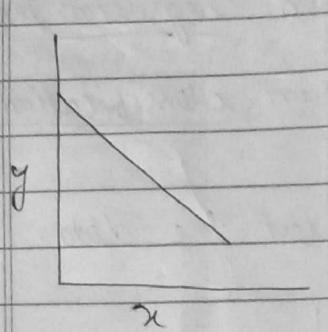
twitter 

Telegram 

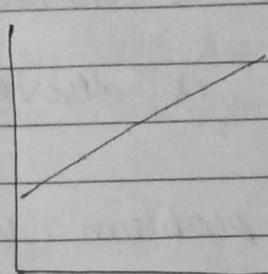
Date _____ / _____ / _____

Saath

Examples of Linear Relationship:-

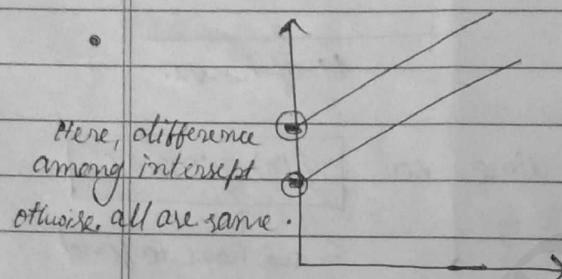


Slope is -ve
neg. Relationship.



Slope is +ve
+ve Relation

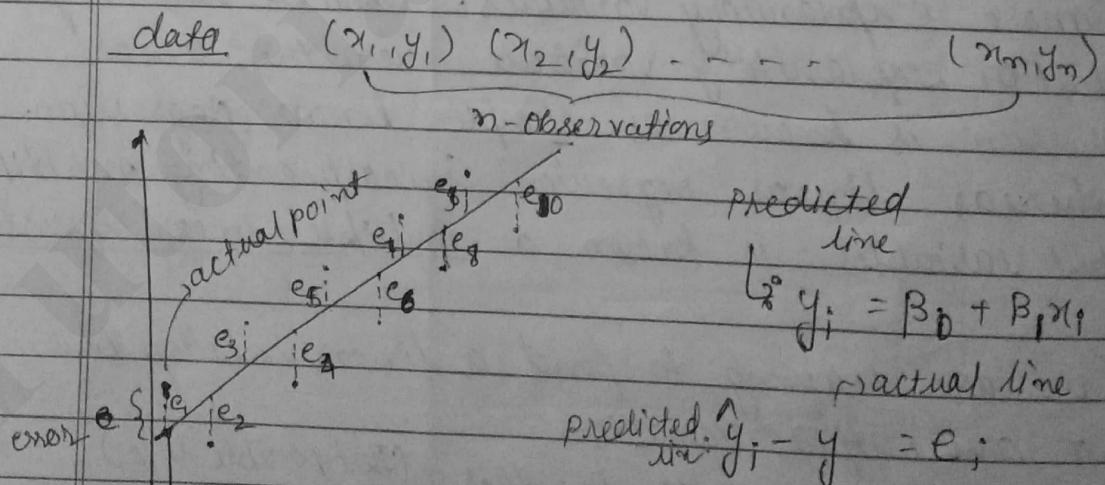
Slope = 0
constant Reln



$$y = c + mx$$

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

Goal: To find β_0 & β_1 such that resulting line is close to these lines.



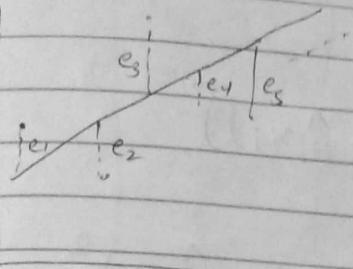
Residual sum of square
 $= e_1^2 + e_2^2 + \dots + e_n^2$

Goal: to minimize RSS,

16+ M subjects
Date: / /

Saathi

Derivation



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

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

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

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

To find minimum.

diff w.r.t β_0 :-

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

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

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

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

Date _____

Saath
Soham

diff w.r.t β_1 .

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

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

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

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

Now. diff

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

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

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

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

Q. Given the data $D = \{(1, 2), (2, 3), (4, 6)\}$

Predict a value for $x=3$. Assume that a simple linear model.

$$\bar{x} = \frac{1+2+3}{3} = \bar{x}_B$$

$$\bar{y} = \frac{1}{3}$$

Saathi

Date _____

i	(x _i - \bar{x})	(y _i - \bar{y})	(x _i - \bar{x}) ²	(x _i - \bar{x})(y _i - \bar{y})
1.	-1/3	-5/3	1/9	10/9
2.	-1/3	-2/3	1/9	2/9
3.	5/3	7/3	25/9	35/9

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

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

or, exp. $\hat{y}_i = \beta_0 + \beta_1 x_i$

$$\hat{y}_3 = \frac{1}{2} + \frac{57}{42} \times \frac{8}{14}$$

Here, x_i = 3

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

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 = $\frac{40}{5} = 8$

$$\beta_1 = \frac{40}{10} = \frac{8}{10} = 0.8 \quad \beta_0 = \frac{14}{5} - 0.8 \times 3 = \frac{14}{5} - 2.4 = \frac{14}{5} - \frac{24}{10} = \frac{14}{5} - \frac{10}{5} = \frac{4}{5} = 0.8$$

$$y_i = 0.4 + 0.8 x_i$$

$$= \frac{14}{5} - \frac{24}{10} = \frac{14}{10} - \frac{24}{10} = \frac{-10}{10} = -1$$

Page No. _____

Saath

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

$$e_1 + e_2 + e_3 + e_4 + e_5$$

$$= (-0.2) + (1) + (-0.6) + (-0.8) + (0.6)$$

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

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

$$= 2.4$$

$$\boxed{RSS = 2.4}$$

Another formula for slope (β_1) :-

$$\beta_1 = \frac{\text{co-variance } (\bar{x}, \bar{y})}{\text{variance } (\bar{x})}$$

we are regressing
y on x.

where, equation of line is :-

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

x is the predicting variable

Variance of input value is as
denomination

Date 23, 01, 19

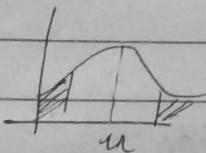
Saathi

Rolling an unbiased dice:-

1,	2,	3,	4,	5,	6	P(x)
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	-	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

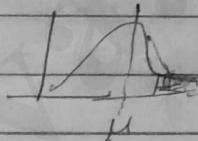
Uniform Distribution (continuous)

$$\bar{x} = u \quad \left\{ \begin{array}{l} \text{2 tailed test} \\ \bar{x} \neq u \end{array} \right.$$



$$H_0: \beta_1 = 0$$

$$\bar{x} > u \quad \left\{ \begin{array}{l} \text{1 tailed test} \\ \bar{x} < u \end{array} \right.$$



$$H_0: \beta_1 \neq 0$$

① Confidence Interval approach

Sample estimate $\pm t\text{-multiplier} \times \text{st}$

$\alpha = 95\%$

dependence

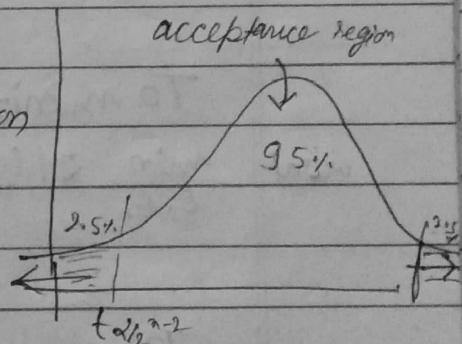
-Q

-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.

Probability value
shaded - rejection region
Non-shaded - acceptance region

→ If H_0 Null hypothesis is true i.e., it lies on Reject region

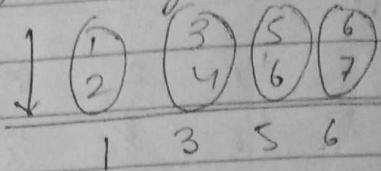
* Less the p-value, evidence against Null hypothesis is more significance.

Date 20/01/19

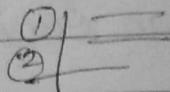
Saath

Lab

$\min(a, \text{axis}=0)$ along x -axis



if axis 1 is along y -axis



31/01/19

31/01/19

In another terminology :- estimated slope

$$h_0(x) = \theta_0 + \theta_1 x$$

hypoth. est. Intercept

↓
est. Intercept

①

Average error

$$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m (h_0(x^i) - y^i)^2$$

↑ ↑
Hypothesis value original value

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

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

Goal

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

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

To simplify ①

$$h_0(x) = \theta_1 x \quad [\text{Intercept}=0]$$

$$\text{10. } J(\theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_0(x^i) - y^i)^2$$

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

Date _____ / _____ / _____

Saathi

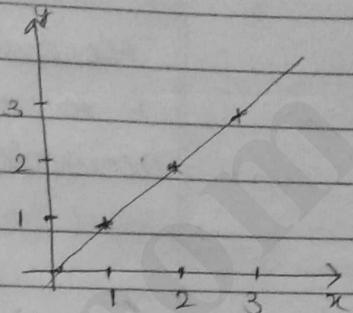
objective $\min J(\theta_1)$

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

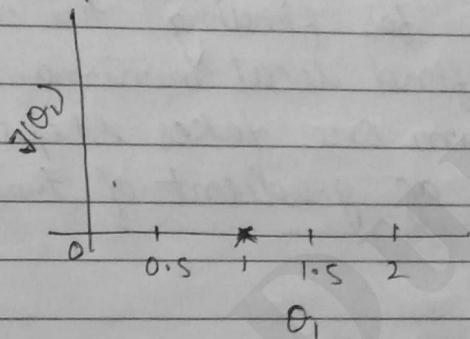
$$J(\theta_1 = 1)$$

$$= \frac{1}{2 \times 3} (0^2 + 0^2 + 0^2)$$

$$= 0$$



Graph of cost function

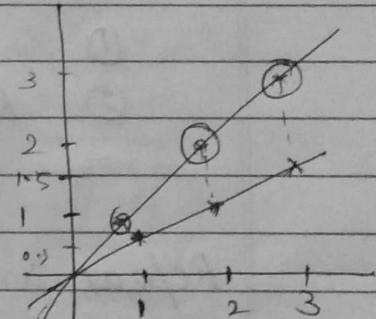


$$J(\theta_1 = 0.5)$$

$$[y = 0.5x]$$

$$= \frac{1}{2 \times 3} [(0.5)^2 + 1^2 + (1.5)^2]$$

$$= 0.58$$

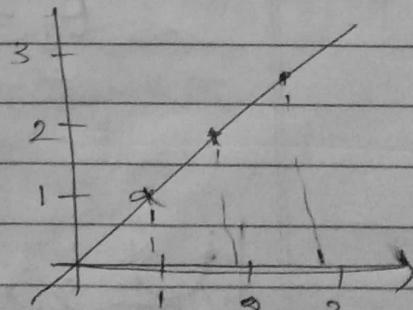


$$J(\theta_1 = 0)$$

$$[y = 0]$$

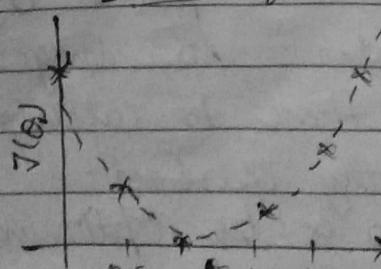
$$= \frac{1}{2 \times 3} [1^2 + 2^2 + 3^2]$$

$$= \frac{1}{3} = 2.3$$



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 make actual line.

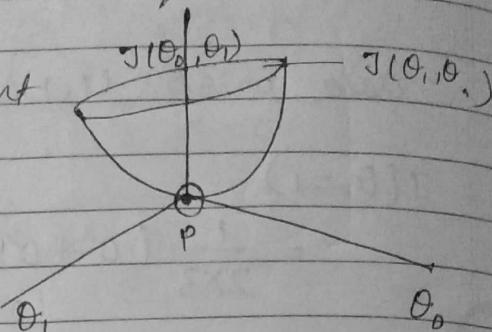


Date _____

Saath

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 .

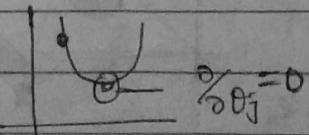
② keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$ until we converge.

Repeat until convergence

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

learning rate.

slope of the curve at θ_j
Here we moving in -ve direction
of slope.



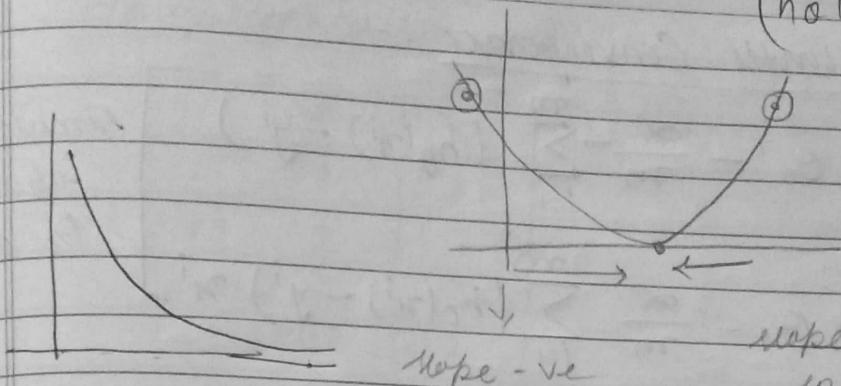
so at this point $\theta_j = \theta_j$

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 , then)
 $(h_\theta(x_i) - y_i)$ Saath



slope +ve

so θ_1 decreased and
move forward

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

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

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

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

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

differentiate w.r.t to θ_1 .

$\underline{\theta_0 + \theta_1 x^i}$
 $(h_\theta(x^i))$ hta

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

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

$$\text{grad} = [\underline{\omega}, \underline{\alpha}]$$

Saath!

Date _____

Repeat until Convergence :-

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

simultaneous update
 $\theta_0 \& \theta_1$

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

3

Simultaneous update

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

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

then update the whole value by

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

$\theta_1 = \theta_1 - \text{tempo}$ so that we don't 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 for i-th example

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

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

Page No. _____

Date _____

Saathi

First Iteration

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

Follow

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

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

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

$$\theta_1 = 0 - \frac{0.1}{5} [(-1)x_1 + (-3)x_2 + (-3)x_4 + (-2)x_3 + (-5)x_5]$$

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

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

Second Iteration

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

$\theta_0 = 0.28$	$\theta_1 = 0.0$	X	Y	$h_{\theta}(x^i) = \theta_0 + \theta_1 x$
1	1			$\frac{7}{25} + 1 \cdot 1 = \frac{32}{25}$
2	3			$\frac{7}{25} + 2 \cdot 1 = \frac{57}{25}$
4	3			$\frac{7}{25} + 4 \cdot 1 = \frac{107}{25}$
3	2			$\frac{7}{25} + 3 \cdot 1 = \frac{82}{25}$
5	5			$\frac{7}{25} + 5 \cdot 1 = \frac{132}{25}$

Date _____

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

$$25) 5.6(2.24 = \frac{7}{25} - \frac{0.1}{5} \times \frac{70}{25}$$

$$\begin{matrix} 50 \\ \cancel{50} \\ \cancel{100} \end{matrix} = \frac{7 - 1.4}{25} = \frac{5.6}{25} = 0.232$$

$$= 0.23$$

$$\theta_1 = 1 - \frac{0.1}{5} \left[\frac{7}{25} x_1 - \frac{18}{25} x_2 + \frac{32}{25} x_4 + \frac{32}{25} x_3 + \frac{7}{25} x_5 \right]$$

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

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

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

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

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

$J(\theta) < \text{threshold value.}$
 θ 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

facebook

WhatsApp 

twitter 

Telegram 

Date _____

Saath

Date _____

→ Multiple Regression :-

Feature				(Y) → Response
x_1	x_2	x_3	x_4	
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

$x_j^i \leftarrow j^{th}$ feature of i^{th} example.

$$X^2 = \begin{bmatrix} 1416 \\ 3 \\ 2 \\ 40 \\ 9 \end{bmatrix} \quad X_2^2 = 3$$

- m is no. of example

- m number of feature.

$$h_{\theta} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

If we incorporate $x_0 = 1$

$$h_{\theta} = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_n x_n$$

~~it may~~ And as.

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

Date / /

Saathi

after incorporating x_0 we get

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

so

$$h_0 = \theta^T x$$

Transpose of θ .

$$J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2m} \sum (h_0(x^i) - y^i)^2$$

Updated Gradient descent formulae:-

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

ie.

$$\theta_0 = \theta_0 - \frac{\alpha}{m} (h_0(x^i) - y^i) \cdot x_0^i [x_0^i = 1]$$

and

$$\theta_1 = \theta_1 - \frac{\alpha}{m} (h_0(x^i) - y^i) x_1^i$$

So general updated formula is

$$\theta_j = \theta_j - \frac{\alpha}{m} (h_0(x^i) - y^i) x_j^i$$

and $x_0^i = 1$ for all x_0^i .

Date 5, 02, 19

Saath

Normal Equation Method

x_0 : remains always 1	x_1	x_2	x_3	x_n	y
1, so feature	x_0	x_1	x_2	x_3	x_n
starts from x_1 ,	1	-	-	-	-
	1	-	-	-	-
m	1	-	-	-	-
	1	-	-	-	-
	1	-	-	-	-
	1	-	-	-	-

X y

$n+1$ features

m = examples.

so, From Data Set.

$[X]$

$m \times (m+1) \rightarrow$ as. $n+1$ feature including

$[Y] m \times 1$

Normal Equation is :-

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

Dimensions:

- $(n+1) \times n$ (number of columns)
- $m \times (m+1)$ (number of rows)
- $(m+1)(m+1)$ (number of columns)
- $(m+1) \times m$ (number of rows)
- $(n+1) \times m$ (number of columns)
- $m \times 1$ (number of rows)
- $(n+1) \times 1$ (number of columns)

Date _____

Sarthi

Gradient descent

1. Iterative

2. Can't work with large value of n

Normal equation

1. Non-iterative

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

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

$$\text{Range} = 45 - 12 = 33,$$

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

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

$$\approx -0.545, 0.121, 0.454, -0.454, 0.30, 0.121$$

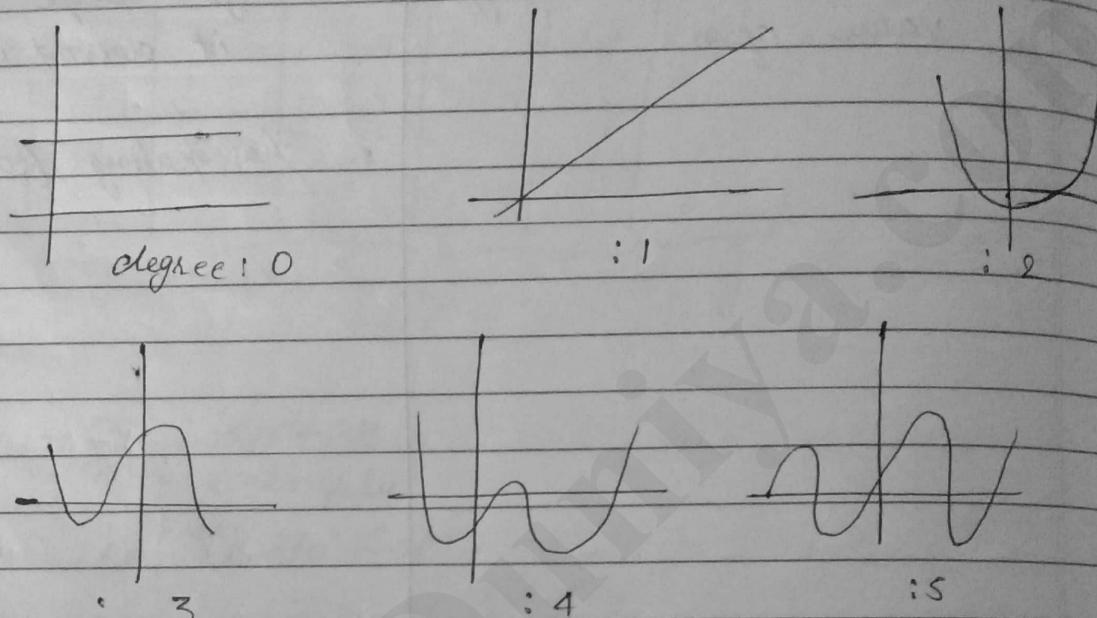
Page No. _____

CH-6 :- page 203 - 209
Date _____

Saathi

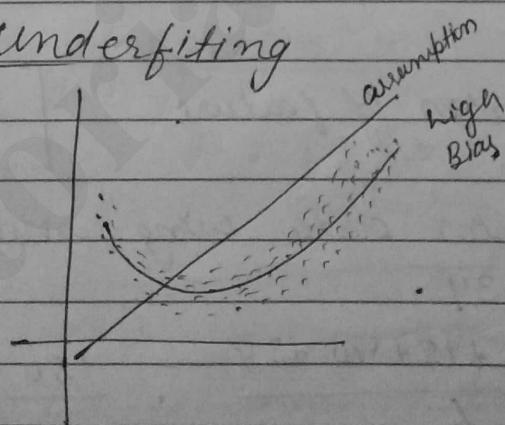
Model :- is a system that maps input to output.

degree of polynomial



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

Underfitting



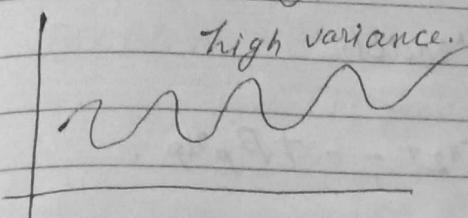
If we assume that it is linear. So we represent it as a linear expression. So our model is not able to find the relationship

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

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

Date _____ / Overshooting

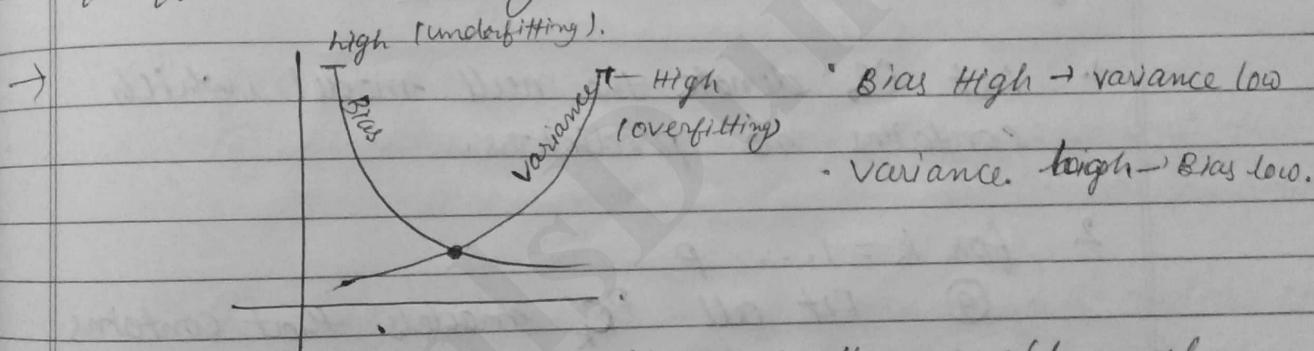
Saathi



- It lost its predictable capability.
- It generates error during test data. Because it has more flexibility.

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

Variance It refers 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 were use.



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

12/02/19

Prediction Accuracy

- # of examples
 - True linear \rightarrow low bias
 - $n \gg p$ \rightarrow low variance
 - $n > p$ \rightarrow high variance
 - $n < p$ \rightarrow infinite variance

Model Interpretability

Date _____

Saathi

Standard Linear Model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Least Square Approach

3 Methods to improve the performance of Least Square Approach

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 $k = 1, \dots, p$
 - (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, \dots, M_p using cross validated prediction error, AIC, BIC or adjusted R^2 .
- R^2 (Coefficient of determination)
Higher value & close to 1, is more accurate
range of R^2 is $[0 < R^2 < 1]$

Page No. _____

Date _____

Saathi

Problem :

(1)

Computational problem. (Computationally Infeasible)

$$P_{C_0} + P_{C_1} + P_{C_2} + \dots + P_{C_P}$$

$$= 2^P.$$

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

(2)

Algo-2 Backward Selection (widely used)

- 1.) Let M_0 denote the null model, which contains no predictors.
- 2) for $k = 0, \dots, p-1$
 - 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 $M_0 - M_p$ using cross validated prediction error, AIC, BIC or adjusted R^2 .

# of features	# of model
$M_0 \Sigma 1$	$P-0$
$M_1 \Sigma 2$	$P-1$
$M_2 \Sigma 3$	$P-2$
\vdots	\vdots
$M_{P-1} \Sigma 1, \dots, P-1$	1
M_P	

as algo-1, algo-2 (subset selection work for least square - so it won't work for me).
Date _____ / _____ / _____

Saathi

Problem 11

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

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

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

$$\text{Algo-2}. \quad 1 + \frac{P(P+1)}{2} = 1 + \frac{20 \times 21}{2}$$

$$= 211.$$

which is far less value.

Algo-3 Backward Subset Selection

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

2) for $k = P, P-1, \dots, 1$.

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

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

3) Select single best Model $M_0 - M_P$.

Date _____ / _____ / _____

Saathi

	# of feature	# of models
M _p	$\{1 \dots p\}$	P-1
M _{p-1}	$\{1 \dots p-1\}$	P-2
⋮		
M ₁	{1}	0
M ₀		1

as, for $n < p$ least square approach can't work.

ie, Here in Backward Subet 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)

print(boston.target)

- n feature

- y set

Saathij

Date _____ / _____ / _____

1. From pandas import *
df = read_csv ("civvers")
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 :-
Continuous 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 of
output.

Date _____ / _____ / _____

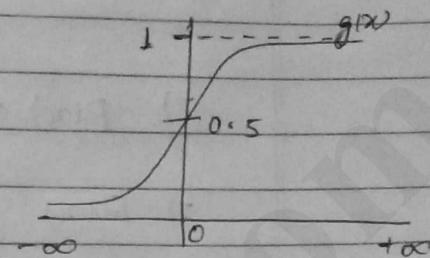
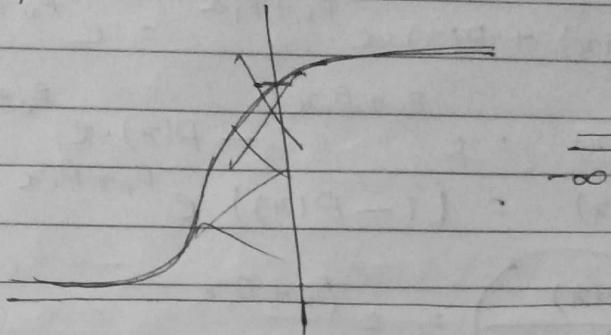
SaathE

Logistic function
(sigmoid function).

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

Output : [0, 1]

graph



$P(x) = \beta_0 + \beta_1 x$ it gives ~~set~~ output as any number (+ve, -ve) so we map this 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.

algo

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

↳

In Linear Regression we have

$$h_{\theta}(x) = \beta_0 + \beta_1 x \quad \text{so, Probability } P(x)$$

$$P(x) = \beta_0 + \beta_1 x$$

and Logistic Regression.

Probability of (P(x))
belonging to class

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

It may go beyond 0-1 range so we apply logistic function.

Estimation for β_0 & β_1

• Maximum-likelihood Estimation.

Saath

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 \leftarrow

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

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 than 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

response is default = yes

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

Date _____

(Saathi)

For Balance : \$1000.

(Table Form ①)

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

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

$$= 0.00576 < 0.5$$

which is less than 1%. so

not a defaulter.

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

a. For Table ②

$$P(\text{default} = \text{yes} / \text{student} = \text{yes})$$

$$= 3.5041 + 0.4049 \times 1$$

as student = yes
 $x = 1$

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

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

facebook

WhatsApp 

twitter 

Telegram 

Lab

Date 20/02/19.

(Saath)

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

Table

Simple logistic	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

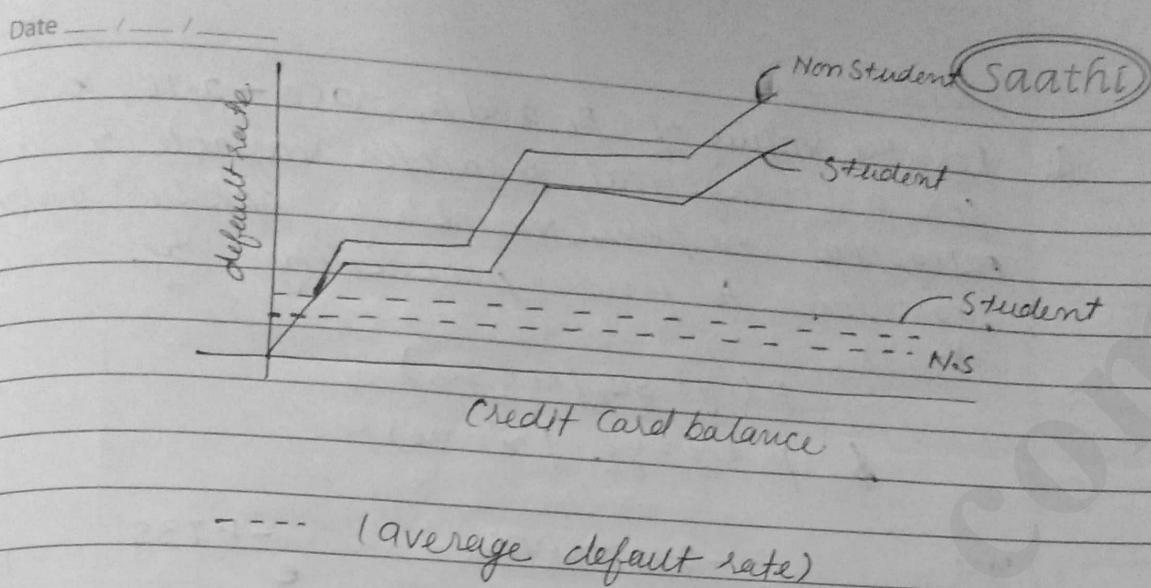
$$\log \left(\frac{P(x)}{1-P(x)} \right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

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

Multiple logistic

	coeff	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	-9.74	0.0002

Page No.



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.

- Predict the probability of default for student and non-student with a credit card balance \$1500 and income \$ 40,000.
- $P(\text{default} | \text{yes, student} = \text{yes})$.

$$\begin{aligned}
 P &= \frac{e^{-10.8690 + 0.0057 \times 1500 + 0.0030 \times 40 + (-0.6468) \times 1}}{1 + e^{-10.8690 + 0.0057 \times 1500 + 0.0030 \times 40 + (-0.6468) \times 1}} \\
 &= \frac{e^{-10.8690 + 8.55 + 0.012 - 0.6468}}{1 + e^{-10.8690 + 8.55 + 0.012 - 0.6468}} \\
 &= \frac{e^{-2.9538}}{1 + e^{-2.9538}} = \boxed{\frac{-0.2355}{1 - 0.2355} = 0.1645}
 \end{aligned}$$

Saath

Date / /

- Q. For the values of β_0 and β_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

$$P(y = \text{yes} | x = \text{yes})$$

$$\$ P(y = \text{yes} | x = \text{no})$$

$$x=1 \quad -2.16 + 0.425 \quad -1.735$$

$$P_1 = \frac{e}{1+e^{-2.16+0.425}} = \frac{e}{1+e^{-1.735}} = \frac{e}{1+0.176} = 0.883$$

$$= \frac{0.883}{1+0.883} = \frac{0.883}{1.883} = 0.487 = \frac{49}{99}$$

$$P_2 = \frac{e^{-2.16}}{1+e^{-2.16}} = \frac{0.515}{1+0.515} = \frac{0.515}{1.515} = 0.333$$

$$= \frac{0.333}{1+0.333} = \frac{0.333}{1.333} = 0.25 = \frac{25}{100} = 0.25$$

22/02/19 Logistic Regression

$$\text{goal} \rightarrow 0 \leq h_{\theta}(x) \leq 1$$

where $h_{\theta}(x) = \theta^T x$ in case of linear Regress

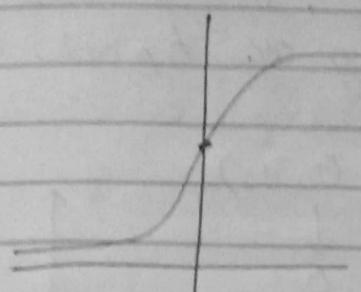
$$\text{Here } \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \quad x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

For Logistic Regression.

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

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

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



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

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

$h_{\theta}(x) = 0.7$

[belongs to class 1]

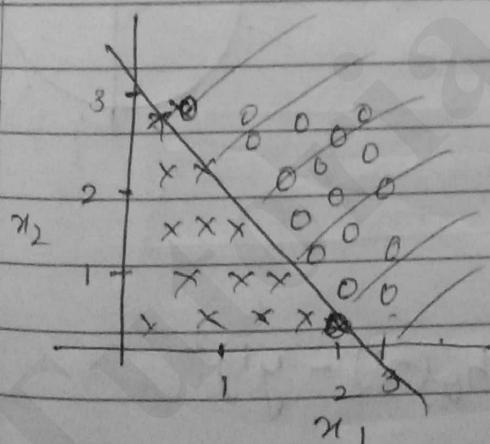
Probability of belonging to first class $P(y=1/x; \theta)$ (class 1)
 $P(y=1/x; \theta) = 0.7$. class 1 because
 $h_{\theta}(x) > P(y=0/x; \theta)$
i.e. $0.7 > 0.3$.

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

$\underbrace{\hspace{1cm}}_{h_{\theta}(x)}$

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

Decision Boundary



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

Suppose:- $\downarrow \quad \downarrow \quad \downarrow$

$$-3x_1 + x_2 > 0$$

$$x_1 + x_2 > 3$$

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

Saathi

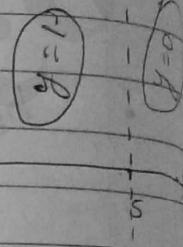
Date _____

- 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 $h_\theta(x)$.

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

$$\begin{matrix} h_\theta(x) = & g(5 - x_1) \\ & 5 - x_1 > 0 \\ & 5 > x_1 \end{matrix}$$

[For positive input
bcuz only pos input
we have outcome
as greater than 0.5]



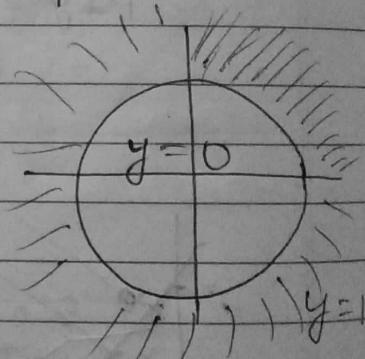
Decision boundary can be non-linear.

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

$$\theta^T x > 0$$

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

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



Cost Function

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

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

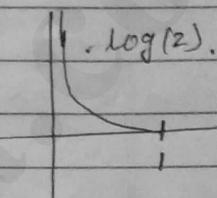
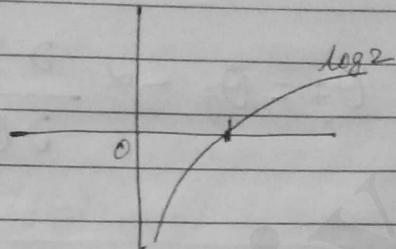
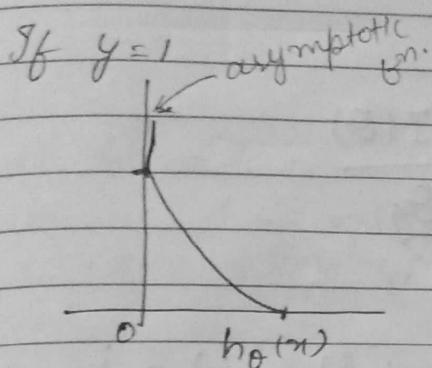
$$= \frac{1}{2} \sum_{i=0}^m \text{cost}(h_\theta(x^i), y^i)$$

3
Saath

Date _____

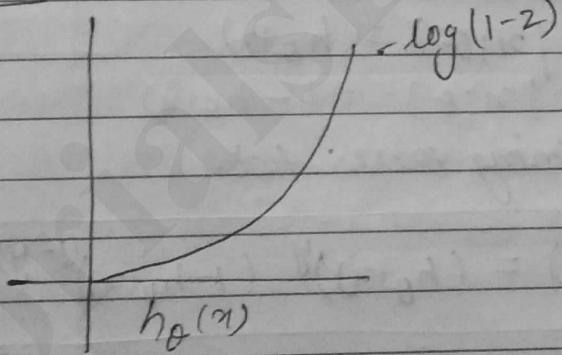
In case of logistic function we take cost function as -

$$\text{cost}(h_0(z), y) = \begin{cases} -\log(h_0(z)) & \text{if } y=1 \\ -\log(1-h_0(z)) & \text{if } y=0. \end{cases}$$



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

For $y=0$



$$\boxed{\text{cost}(h_0(z), y) = -y \log(h_0(z)) + (1-y) \log(1-h_0(z))}$$

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

when $y=0$

$$\text{cost} = -\log(1-h_0(z))$$

and

$y=1$

$$\text{cost} = -\log(h_0(z)).$$

Date / /

Saath

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

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

To find θ 's

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

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

#

Maximum Likelihood Approximation:-

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

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

On combining these two.

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

Likelihood

$$L = P(y|x; \theta)$$

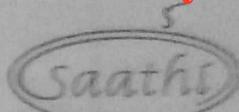
$$= \prod_m P(y^i|x^i; \theta)$$

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

For mathematical simplicity we convert it in log function.

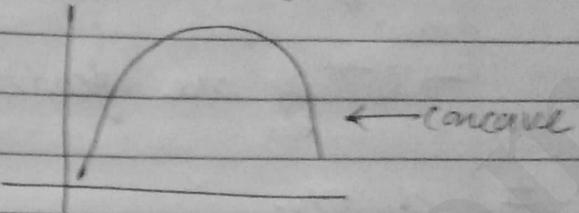
26/0

Date _____



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

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



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

i.e.

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

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 maximizes the likelihood function.

In practice it is actually 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
- you do not need to choose the value of α .
 - They are faster than gradient descent.

disadv.

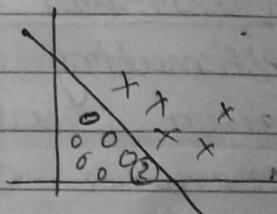
They are more complex as compared to gradient descent.

Multiclass Classification

One Vs. All approach
or

One Vs. Rest.

If we have binary class
so we can do binary classification

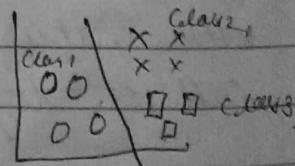


But if more than two categories then we can't follow the above approach.

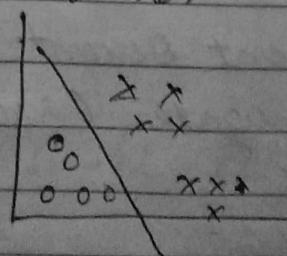
One vs all or Rest

For checking class 1.

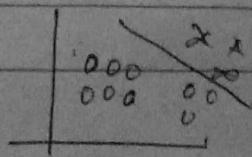
we just find the decision boundary for class 1. That is whether the particular point belongs to class 1 or not.



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

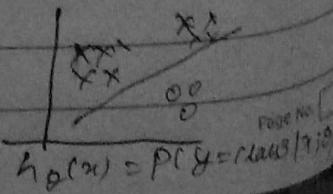


For class 2



For class 3

$$h_0^1(x) = P(y=\text{class 2} | x; \theta)$$



$$h_0^3(x) = P(y=\text{class 3} | x; \theta)$$

Date _____

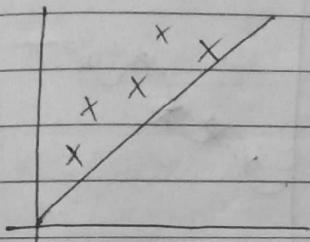
Saathi

Find.

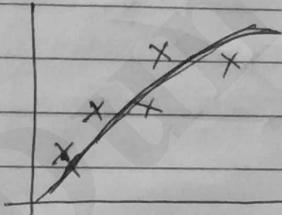
$$\max_i h_{(0)}^i(x)$$

- 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 end up training.
- i) $K-1$
 - ii) K
 - ✓ iii) $K+1$
 - iv) $\log(K)$

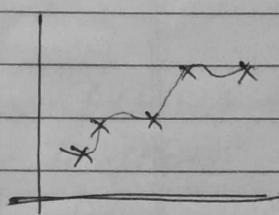
Regularization



- underfitting
- Model oversimplifies 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 variance.
- due to lots of feature than no. of example locations
- High degree polynomial

- Q. Consider a medical diagnosis of classifying tumors as malignant or benign. If a hypothesis $h_{(0)}$ has ~~overfit~~ the learning set. It means when one of the following.

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

Saath

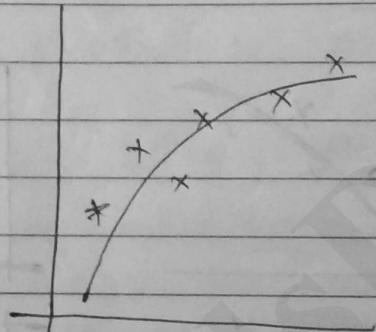
Date _____ / _____ / _____

⑤ It doesn't make accurate for example in the training set, but it generalizes well on new examples.

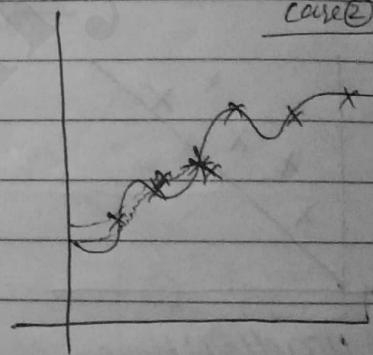
⑥ It makes accurate prediction for examples in training set but not generalizes well on new eg.

⑦ It doesn't make accurate predictions for examples in training set. and doesn't generalise to make prediction on new examples

overfitting.
cause ②



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



$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3^3 + \theta_4 x_4^4$$

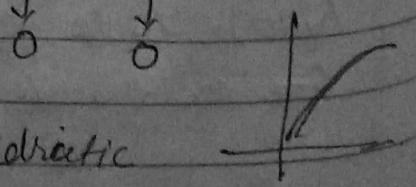
With regularization we are trying to decrease the flexibility. So that reduce the variance and increase bias.

But if Take all attributes but reduce the coefficient of each attribute which leads to normal form.

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

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

↳ Quadratic



Date _____

Saathi

In case of overfitting:- case(2)

$$\min J(\theta) = \frac{1}{2m} \sum_{j=1}^m (h_{\theta}(x^{(j)}) - y^{(j)})^2 + 1000x_3^3 + 1000y_1^2$$

If it have higher coefficient so it also need to reduce at the 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 \rightarrow 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

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

Regularization term
Regularization co-efficient

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

- 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 overfitting.

Saath

Date _____

- vii) Algo results an underfitting.
- iv) Gradient descent will fail to converge.
- iii) because to make the term we need to take value 0.000000... thus, we'll be left with only θ_0 , that it'll only make a line and all points will not be covered. Thus, underfitting.

Regularized Gradient descent - will also change as cost funcn also changed.

repeat {

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

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

Taking θ_j common :-

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

\downarrow
first term λ

thus θ is shrinking.

Normal Equation Method

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

For Regularization,

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

$(n+1) \times (n+1)$
 \downarrow
no. of features.

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

Regularized Logistic Regression

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

Gradient descent
repeat {

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

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

{}

14/03/19

Confusion Matrix

↳ A confusion matrix is a table often used 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 negative - actual out +ve & Predictive -ve

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

facebook

WhatsApp 

twitter 

Telegram 

4 Cb

Date _____

		Predicted	
		T	F
actual.	T	TP	FN
	F	FP	FN

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

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

Precision \rightarrow 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 predict a class variable as yes or no. The class variable could predict correct yes for 230 no. of rows, ~~incorrect~~ correct 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 misclassification error.

Say n:

- Accuracy = $\frac{230+120}{500} = \frac{350}{500}$
 $= \frac{7}{10} = 0.7$

Predicted

		T	F
Actual	T	230	50
	F	100	120

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

Date _____ / _____ / _____

Saath

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

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

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

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

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

$$= \overset{\downarrow}{g(z)} (1-g(z))$$

So,

$$g'(z) = g(z)(1-g(z))$$

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

H.M Bipolar sigmoidal function and its derivatives

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

Neural Networks

Date 29, 03, 19

Saathi

Q Diff biological Neuron & Artificial Neural Networks.

• Artificial Neural Networks are information built from interconnected elementary 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 analogous to neuron in the human brain. Each node has no. of inputs, associated node function and an output.

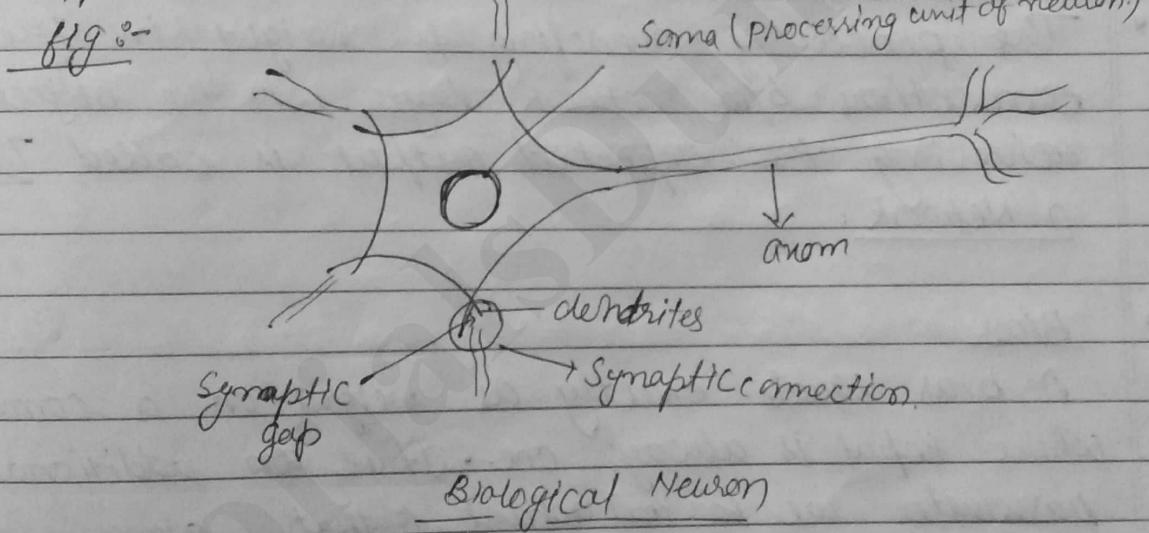
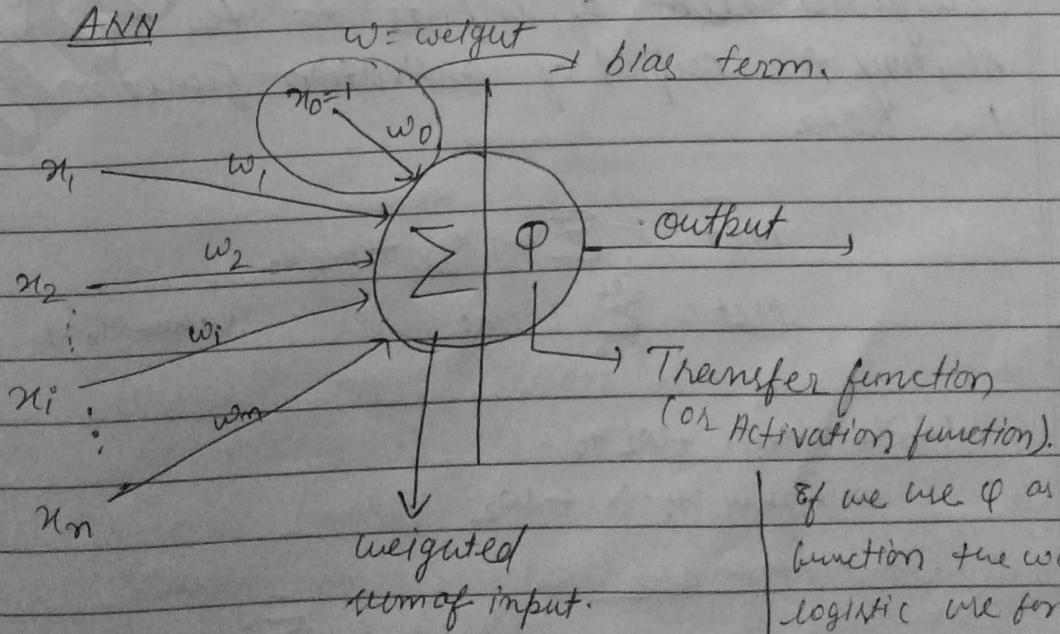


Fig. ANN



weighted sum of input.

If we use Φ as sigmoidal function the work as logistic use for classification.

Date _____ / _____ / _____

• Weight

In a Neural Network, neurons are connected to each other, using connection units. 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 incidenting on a neuron.

$$\boxed{\text{Net} = \sum w_i n_i}$$

The process of modifying 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 we 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.

$$\text{Net} = \sum w_i n_i + w_0 n_0$$

$$\text{Net} = \sum_{i=0}^m w_i n_i \quad \text{when } n_0 = 1$$

$$w_0 n_0 + w_1 n_1 + w_2 n_2 -$$

$$w_1 n_1 + w_2 n_2 > -w_2$$

Let it also as

$$\text{Net} = b + \sum_{i=1}^m n_i w_i$$

Date _____

Saathi

Activation funⁿ.

Activation funⁿ. converts input signal of a node 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.

$$(P(z)) = z$$

① Different types of activation function :-

- Linear Activation function.

$$f(u) = au$$

range = $-\infty$ to ∞

use - suitable to use at O/P layer.

- Sigmoidal function

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

range 0 - 1

use : O/P layer of binary classification problem.

- tanh - mathematically shifted version of sigmoidal function.

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

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 input layer.

Date _____ / _____ / _____

$$\text{sgn}(n) = \begin{cases} 1 & \text{if } n > 0 \\ -1 & \text{otherwise.} \end{cases}$$

Diff

ANN

- 1.) Processing speed $\sim 10^{-18}$
- Parallelism = comparatively less II.
- size & complexity don't involve this much computational.
- New info destroys the old info.
- individual unit in ANN o/p a single const. value.

Biological Network N/w

- Processing speed 10^{-3} s
- It is massively II architecture
- Estimated to contain 10^{11} neurons each connects to 10^4 other neurons
- New info doesn't destroy the old info.
- 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, discrete 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, maybe required.
6. The ability of humans to understand the learned target function is not important

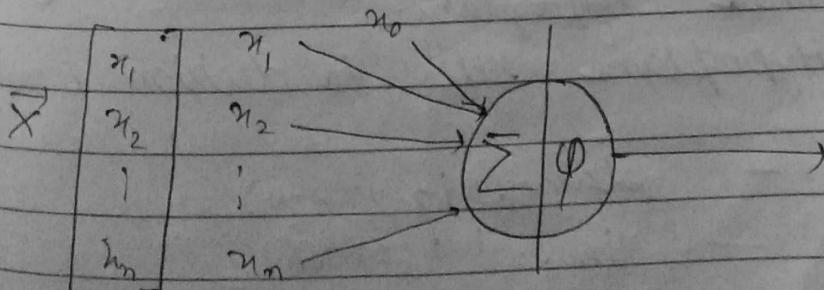
Perception

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.

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

also called sign.

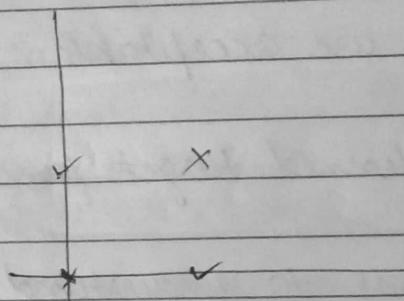
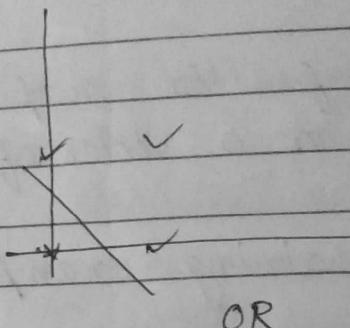
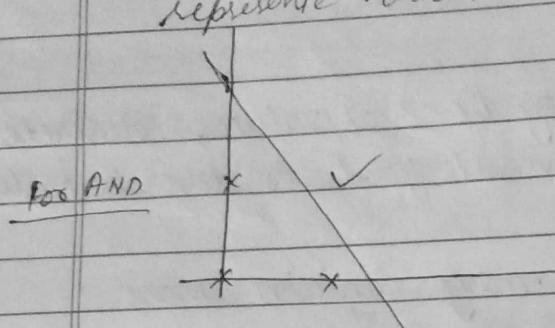
(output is 1 or -1)



Date — / — / —

Q. Architecture of Perception?

Problem associated with single layered Perception:
represents various boolean funct.



I can't make separate these
with help of only one decision
line.

Perception won't be able

Q. Define the concept of linear separability / problem with perf.
show -- (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. ($2-1=1-D$).

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 until the perceptron classifies all training example correctly.

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

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

↓ target out. ↓ current output. → input

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 _____

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

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

$$o_d = \bar{w} \cdot x_d$$

mean comb.

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

$$= - \sum_{d \in D} (t_d - \bar{w} \cdot x_d) x_i d$$

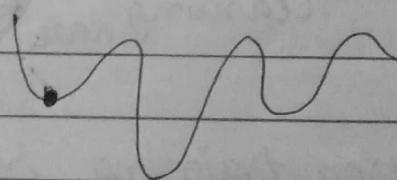
On substituting

$$\Delta w = \eta \sum_{d \in D} (t_d - \bar{w} \cdot x_d) x_i d$$

: delta Rule.

problem

- This process might be slow. (overcome by stochastic plasticity at local minima.)



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

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

Standard Gradient descent Stochastic Gradient descent

1. Error is summed over all the examples before updating the weights

1. Weights are updated upon examining each training example

Date _____

Saathi

- | | |
|--|--|
| 3. 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 has 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 are estimated for every observation as oppose to whole samples. In, It reduces a lot of random ness. the path of this randomness over more places and thus avoid getting into local minima. |
| 5. $\Delta w_i = n \sum_{d \in D} (t_d - O_d) x_{id}$ | 5. $\Delta w_i = n(t - O)x_i$ |

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

Perceptron rule

A rule

- | | |
|---|---|
| 1. This rule updates the weight based on output from thresholded perceptron | 1. This rule updates the weight based on unthresholded perception |
| 2. output: | 2. output |

Slices

Date _____ / _____ / _____

Saathi

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

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

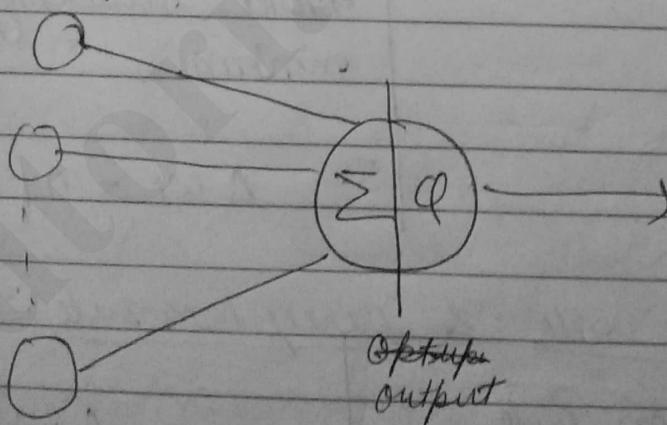
4.

③ Linear Programming :-

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

→ Use only for linearly separable type examples.

• Single layer

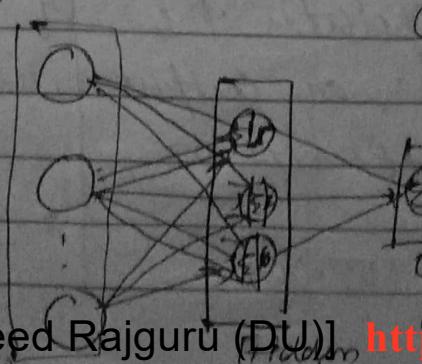


Input
layer.

• Multi layer

- No reverse link
(no called feed forward
network)

bt error are backpropagation
in multi layer



Q. Draw a diagram
for multilayer
perceptron

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

A unit in multilayer architecture should have -

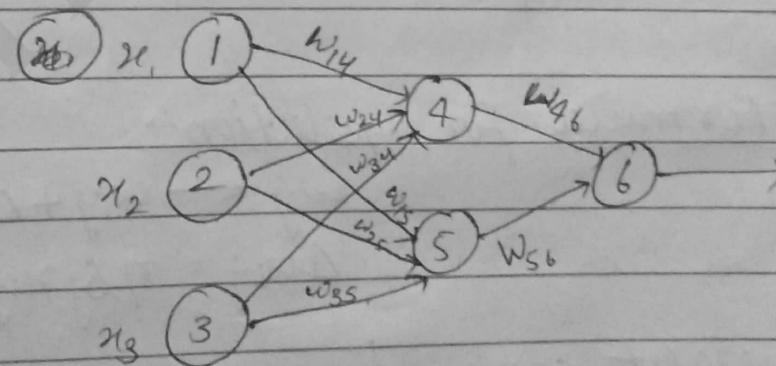
1. Output is non-linear function of inputs.
2. Output is differentiable function of input.

9/04/19

Q. Consider a multi-layer feedforward 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 $x = (0, 1)$ with class label 1.

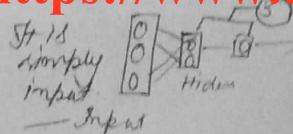
- 1.) Calculate net input and output of each neuron.
- 2.) Calculate error at each node.
- 3.) Calculate updated weight and bias.



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

$\theta_0 = -2.0$ (w_0)
also $\theta_0 = 1$ θ_0

No. of neuron = 3.



(Saath)

Date _____ / _____ / _____

Solution: ①

Unit
4

Net Input

$$0.2 + 0 - 0.5 + (-0.4) \xrightarrow{\text{bias}} \frac{0.2 - 0.5 - 0.4}{\theta_0} \text{ here } \theta_0 = 1 \\ = -0.7$$

$$\text{O/P.} \frac{1}{1+e^{-0.7}} = 0.332$$

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

$$0.474(1 - 0.474)(1 - 0.474) \\ = 0.1311$$

$$\delta_k = \partial_k(1 - \partial_k)(t_k - \partial_k)$$

• For hidden layer

5.

$$0.525(1 - 0.525) * (-0.2) \times 0.1311 \\ = -0.0065$$

$$\delta_h = \partial_h(1 - \partial_h) \sum_{k} w_{hk} \delta_k$$

Keep boundary

4

$$0.332(1 - 0.332) \times (-0.3) \times \\ 0.1311 \\ = -0.0087$$

③

Formula for update:-

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

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

Weight or Bias

New Value

w_{46}

$$-0.3 + 0.9 \times 0.1311 \times 0.332 = 0.21$$

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

Saa

$$\theta_6 \quad 0.1 + 0.9 \times 0.1311 \times 1 = 0.218$$

$$\theta_5 \quad 0.2 + 0.9 \times (-0.0065) \times 1 = 0.194$$

$$\theta_4 \quad -0.4 + 0.9 \times (-0.0087) \times 1 = -0.408$$

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

facebook

WhatsApp 

twitter 

Telegram 