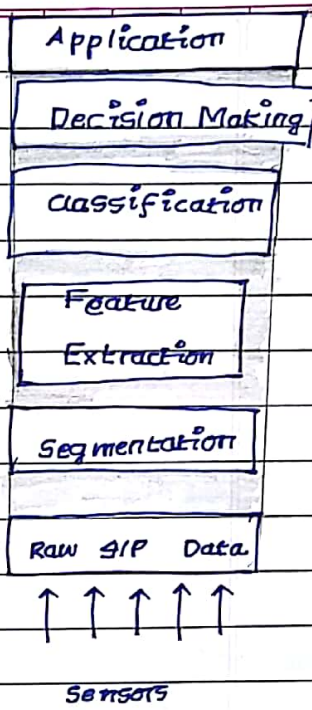
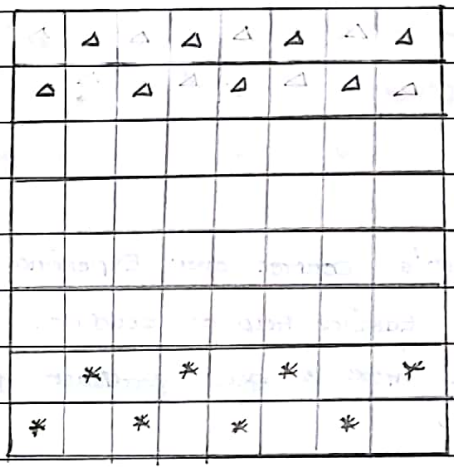
11th August, 2017 Day 1 10:45 - 12:45

DATE: PAGE:

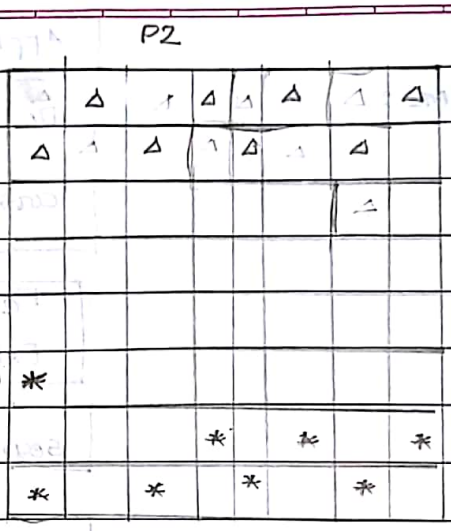
OSI Analogy for ML :



Initial State for Board 'B' :



DATE: PAGE:



Now we are at a "Successor Board State", b

We can have multiple successive state of the board after the first move itself.

P1 could have chosen to move to anywhere else diagonally on 2nd (i=2) row.

Similarly, P2 could have moved to some other position

What is experience?

What we want to learn is all such successive board state corresponding to each move such that as a player I win the game.

Teacher versus learner's control over Experience

- Learning is done taking help of teacher.
- learner makes a move & gets feedback from teacher.
- Sometimes, you play game against oneself & learn.

Direct Training Experience:

Assuming an ideal scenario,

you have data of all possible combination of successive board state per move, from we know at ALL TIMES

DATE: PAGE:

the move we need to make to WIN the game.

We don't have such a data in reality.

Indirect Training Experience:

Choose optimal next move given a current board state, so that you MAY WIN the game. Its not for sure that you'll win the game.

Since the best move is not known, we can assign a value for every board state for move and it reduces to an optimization problem.

We can define a target fn $V(b)$ for an arbitrary board state:

1. If board state b is final + won, $V(b) = +100$
   If b is final + fail, $V(b) = -100$
   } when b is a final state, you can assign such a value.

Together for the FIRST MOVE itself, we could have had another "successive board state" for FIRST MOVE itself.

2. If final & drawn, $V(b) = 0$

3. If b is not final, $V(b) = V(b')$
   where b': best final board state starting from b & playing optimally.
   } we can't just assign a value when b is not a final state

similar to Back Propagation: we know how final state b' looks like & we already have defined V for it (1 & 2), we can recursively assign other values to plausible prior board states.

b': best final state if played optimally.

we need to learn the fn V approximately (i.e., $\hat{V}$)

How to choose a representation for the fn approximation?

We have several possibilities:

we can choose linear,

Quadratic,

Polynomial,

ANN etc.

DATE: PAGE:

In case we choose linear fn,
we'll have to see how much does each of the
following affect the chances of winning?

i) # of black pawns for current board state b : x_1

ii) # " red " " " " " b : x_2

iii) # " black " " " " " b : x_3
that are at threat.

iv) same with red pawns : x_4

So,

$\hat{V}(b)$ can be written as

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4$$

How much does each of
x_1, x_2, x_3, x_4
contribute is determined by trying
w_1, w_2, w_3, w_4.

Chapter 2 DATE: PAGE:

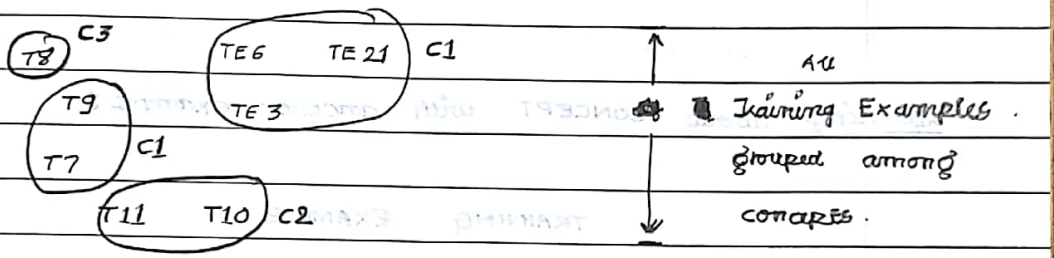
Concept Learning

Ex	beak	wings	legs	fly	Bird (concept)
1	Yes	Yes	Yes	Yes	1 (denoted by a boolean value)
2	Yes	Yes	Yes	No	1 (still a bird denoted by 1)
3	Yes	No	Yes	No	0 (not a " " " 0)

member of concept class

non-member of concept class

$X = \{$ Set of Training Examples $\}$



Domain is set of all training example. $f: \mathbb{R}_{m \times n} \rightarrow$ Boolean $m \times 1$

Concept is a boolean valued fn mathematically.

Otherwise, set of examples that can be grouped together,

or " " instances " " " " " ".

forms a concept

Each instance has set of attributes.

Conjunction and use of concept :

Ex1 from above :

Mathematically,

The object has	beak		$x1$
	AND wings	conjunction ∧	$x2$
	AND legs	∧	$x3$
	AND can fly	∧	$x4$
is a bird.		give	1

OR

$(x1, x2, x3, x4, 1)$

DATE: PAGE:

Ex 2 From abv :

The object has beak — Mathematically, x1
AND wings — ∧ x2
AND legs — ∧ x3
AND cannot fly — ∧ ¬x4
is a bird — gives 1

OR

(x1, x2, x3, ¬x4, 1)

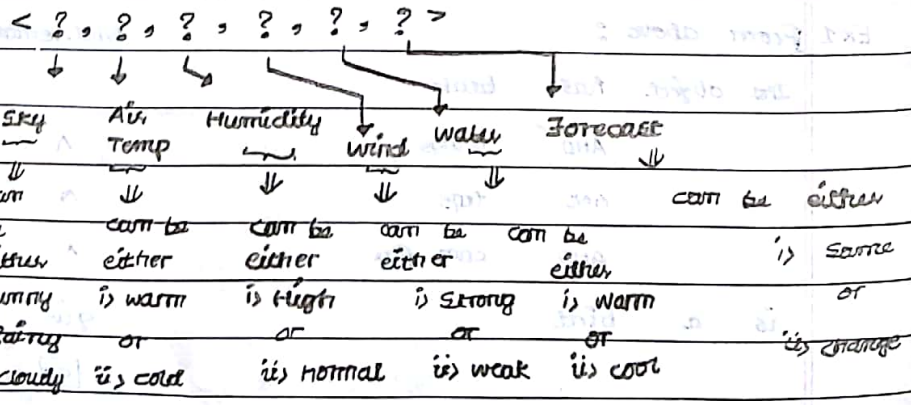
Learning about CONCEPT with another example :

TRAINING EXAMPLE

Day	Sky	AirTemp	Humidity	Wind	Water	Forecast	Sport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	"	"	High	"	"	"	"
3	Rainy	Cold	"	"	"	Change	No
4	Sunny	Warm	"	"	Cool	"	Yes

Whether he plays sport or not is concept.

Most general representation of any instance (or training example) is,



when for an instance, this attribute (here, air temp) value is missing

DATE: PAGE:

$< ?, \phi, ?, \phi, ?, ? >$

Notations :

$X \longrightarrow$ Set of all instances

$c(x) \longrightarrow$ Concept for a given concept.

$h_i(x) \longrightarrow$ The hypothesis for a given instance. This may or may not agree with $c(x)$

Disjunctions can't be represented by our chosen hypothesis.

Each instance is represented as

$< x1, x2, x3, x4, x5, x6 >$

where

it means

if that instance has x1 and x2 and x3 and x4 and x5 and x6

then it may or may not belong to the concept.

Inductive Hypothesis Learning :

DATE: PAGE:

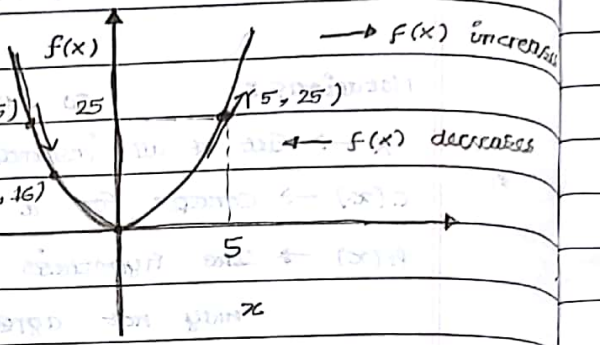
$f(x) = x^2$

$\frac{df}{dx} = 2x$

$\frac{\Delta y}{\Delta x} = \frac{25-16}{-5-(-4)}$

$= \frac{9}{-1}$

$= -9$

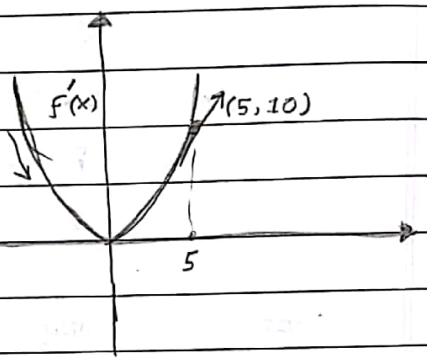


i) @ $x = 5$

$\frac{df}{dx} = 2(5) = 10$

ii) @ $x = -5$

$\frac{df}{dx} = 2(-5) = -10$



For $x = 5$,

we should move in the direction opposite to that of slope.

For $x = -5$,

we should move in the direction of the slope

$y = x$

(1, 1) as $x \uparrow$ $y \uparrow$

$\frac{dy}{dx} = 1$

$|+1| = 2, y = 2$

∴ Slope always points in the direction in which y value increases.

for $y = -x$

1, −1

2, −2

as $x \uparrow$ $y \downarrow$

$\frac{dy}{dx} = -1$.

$2 - 1 = 1$

$y = -1$

$3 - 1 = 2$

$y = -2$

Earlier, $y = -3$

now $y = -2$

∴

22nd August, 2017

$4 \times 3^5 = 4 \times 3^3 \times 3^2$
$= 4 \times 27 \times 9$

DATE: PAGE:

Hypothesis,

Hypothesis for a given instance,

concept " " " "

Suppose,

my hypothesis h_1 = < Sunny, ?, ?, Strong, ?, ? >

If any instance such as

$x1$ = < Sunny, Warm, Normal, Strong, Warm, Same >

$x2$ = < Sunny, Warm, High, Strong, Cold, Same >

which satisfy the constraints of the hypothesis $h1$ irrespective of their concept value will satisfy the hypothesis $h1$.

It means its possible that

For concept C, $x1$ is considered as a Positive example & $x2$ is considered as a -ve example.

$c(x1) = c(<Sunny, Warm, Normal, Strong, Warm, Same>)$
$= <Sunny, Warm, Normal, Strong, Warm, Same, 1>$

where as

$c(x2) = c(<Sunny, Warm, High, Strong, Cold, Same>)$
$= <Sunny, Warm, Normal, Strong, Warm, Same, 0>$

where

c : Plays Game : $X \rightarrow \{0, 1\}$

but both $x1$ & $x2$ falls under or satisfy the (constraints of) hypothesis $h1$ and

$\therefore h(x1) = 1$ [Though, $c(x1) = 1$]

&

$h(x2) = 1$ [" , $c(x2) = 0$]

If we do $2-(-1) = 3$ at $x = 3$, $y = -3$ fn value decreases

DATE: PAGE:

Calculating # of distinct instances :

We can classify instances based on semantic and syntax.

Syntactically,

	Sky	Air Temp	Humidity	Wind	Water	Forecast
	⇩	⇩	⇩	⇩	⇩	⇩
can	Sunny	Warm	Normal	Strong	Warm	Same
take any	Cloudy	Cold	High	Weak	High	Change
of these	Rainy	ϕ	ϕ	ϕ	ϕ	ϕ
values	ϕ	?	?	?	?	?
	?					

Syntactically, # of distinct instances that we can get

$= (3+2) \times (2+2) \times (2+2) \times (2+2) \times (2+2) \times (2+2)$

$= 5 \times 4 \times 4 \times 4 \times 4 \times 4$

$= 5 \times 2^2 \times 2^2 \times 2^2 \times 2^2 \times 2^2$

$= 5 \times 2^{10}$

$= 5 \times 1024$

$= 5120$ syntactically distinct instances

Semantically,

we assume any instance having atleast one ϕ is considered same semantically.

Semantic could be,

if any instance has even one ϕ (that is even if one of the attribute value for that instance is missing)

we'll consider all such instance as redundant.

DATE: PAGE:

For any such instance X_i which is redundant (as per our defined semantic), we do not worry about whether it satisfies any concept C or it " " constraint of any hypothesis h.

Here we defined 1 semantic for all instances having atleast one ϕ (though all such instances are syntactically different) — (1)

Apart from this,

	Sky	Air Temp	Humidity	Wind	Water	Forecast
can take any of these values & each will be semantically different	Sunny	Warm	Normal	Strong	Warm	Same
	Cloudy	Cold	High	Weak	High	Change
	Rainy	?	?	?	?	?
	?					

Semantically distinct instances need not satisfy same concept or hypothesis

Semantically here, # of distinct instances

$= (3+1) \times (2+1) \times (2+1) \times (2+1) \times (2+1) \times (2+1)$

$= 4 \times 3 \times 3 \times 3 \times 3 \times 3$

$= 4 \times 81 \times 3$

$= 4 \times 243$

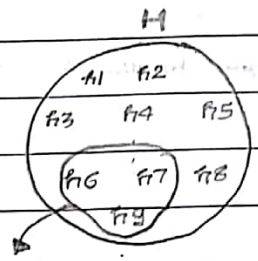
$= 972$ — (2)

Using (1) & (2), total # of semantically distinct instances $= 1 + 972$

$= 973$ semantically distinct instances.

DATE: PAGE:

Version Space



Version Space : Set of hypothesis which satisfies all the training examples in set D such that $h_1(x) = c(x)$

Candidate Elimination Algorithm

1.	Japan	Honda	Blue	1980	Economy	+ve

Starting Point :

So, Most Specific	ϕ	ϕ	ϕ	ϕ	ϕ	
G0, Most General	?	?	?	?	?	

For ex 1 : Since it is a +ve example, start with General side DATE: PAGE:

S0 ϕ ϕ ϕ ϕ ϕ

S1 Japan Honda Blue 1980 Economy — (2) Now find more general compared to S0, specific hypothesis S1

G1, G0 ∵ G0 satisfies ex 1
retain it as G1 — (1)

2. Japan Toyota Blue 1970 Sports -ve

Since ex2 is -ve, start with specific side.

S0 ϕ ϕ ϕ ϕ ϕ

S1 Japan Honda Blue 1980 Economy — (1) ∵ S1 designates ex2 as -ve
∴ Retain it as S2

S2 " " " " "

G2 gives ex2 as +ve
whereas infact it is -ve
∴ we need to ~~accept~~ get a new G2 — (2)

DATE: PAGE:

Getting G2 :

<J, ?, ?, ?, ?> ✗ Not satisfied by ex1 & ex2

<?, H, ?, ?, ?> ✓

<?, ?, B, ?, ?> ✓

<?, ?, ?, 1980, ?> ✓

<?, ?, ?, ?, Eco> ✓

∴ G2 = { <?, H, ?, ?, ?>,
<?, ?, B, ?, ?>,
<?, ?, ?, 1980, ?>,
<?, ?, ?, ?, Eco> }

3. Japan, Toyota, Blue, 1990, Economy +ve.

∵ it is +ve,
start from general side.

Check if ex3 satisfies G2. — (1)

G2 = { <?, H, ?, ?, ?> ✗
<?, ?, B, ?, ?> ✓
<?, ?, ?, 1980, ?> ✗
<?, ?, ?, ?, Eco> ✓ }

∴ G3 = { <?, ?, B, ?, ?>,
<?, ?, ?, ?, Eco> }

(2) Going to specific side we get,

S3 = <Japan, ?, Blue, ?, Economy>

DATE: PAGE:

4. USA, Chrysler, Red, 1980, Economy -ve

ex2 $h(x) \neq c(x)$ for both ex1 & ex2 → $h(x) = c(x)$ / $h(x) \neq c(x)$

S3 : < Japan, ?, Blue, ?, Economy >

Start with specific side

Amounts ex4 as -ve

∴ S3 is retained as S4

S4 : < J, ?, B, ?, E > — (1)

Getting G4 :

G3 → < ?, ?, B, ?, ? > satisfied by ex1, 2 & 3 also satisfies ex4

< ?, ?, ?, ?, Eco > ✗ satisfies by ex1, ex2, ex3 but not by ex4.

∴ Find a more specific hypothesis than G3 (i.e; G4) that satisfies ex1, ex2, ex3, ex4.

We'll revisit from ex1 to ex4 & find out that

< J, ?, ?, ?, Eco > satisfies ex1 to ex4

∴ G4 = { < ?, ?, B, ?, ? >, < J, ?, ?, ?, Eco > } — (2)

DATE: PAGE:

5. Japan Honda White 1980 Economy +ve

Do general first then specific side

G4: { <? ? B ? ?>
<J ? ? ? Eco> }

⑤ x 5 does not satisfy <? ? B ? ?>
but satisfies <J ? ? ? Eco>

∴ Update

G4 as

G5 = { <J ? ? ? Eco> }

Getting S5:

S4: <J ? B ? E>
becomes
S5: <J ? ? ? E>

We stop when
$S_i = G_i$
or
when we run out of training example

$AX = B$
$A^{-1}AX = A^{-1}B$
$IX = A^{-1}B$
$X = A^{-1}B$

Imp for Exam : Rank , Singular Matrix , Eigen value / vectors , Inverse of Matrix , Cofactor , determinant

DATE: PAGE:

Final Version Space we get for

Enjoy Sport Example :

S : { <Sunny , Warm , ? , Strong , ? , ?> } h_1

h_2 h_3 h_4

<Sunny , ? , ? , Strong , ? , ?> <Sunny , Warm , ? ? ? ? ?> <? W ? S ? ?>

h_5 h_6

G : { <S ? ? ? ? ?> <? W ? ? ? ?> }

These are 6 hypothesis in the ~~Hypothesis Space~~ final version space.

Test Data :

A	Sunny	Warm	Normal	Strong	Cool	Change
B	Rainy	Cold	,,	Light	Warm	Same
C	Sunny	Warm	Normal	Light	Warm	Same
D	Sunny	Cold	Normal	Strong	Warm	Same

	h_1	h_2	h_3	h_4	h_5	h_6
A	✓	✓	✓	✓	✓	✓
B	×	×	×	×	×	×
C	×	×	✓	×	✓	✓
D	×	✓	×	×		

x_1 x_2 x_3 $\overline{x_4}$ x_5 $\overline{x_6}$

↓

Play is disjunction of attributes

So : $(x_1 \vee x_2 \vee x_3 \vee x_5) \longrightarrow$ play

↓

disjunction of conjunction

Go : $(\neg (x_4 \vee x_5))$ NOT

DATE: PAGE:

Candidate Elimination Algorithm

Examples :

- Enjoy Sport Example :

Training Examples

Day	Sky	Air Temp	Humidity	Wind	Water	Forecast	Enjoy Sport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes (+ve)
2	Sunny	Warm	High	Strong	warm	Same	yes (+ve)
3	Rainy	cold	High	Strong	warm	Change	No (-ve)
4	Sunny	Warm	High	Strong	cold	Change	Yes (+ve)

Applying Candidate Elimination,

G0 (x) < ϕ ϕ ϕ ϕ ϕ ϕ >

G0 (x) < ? ? ? ? ? ? >

↕ Specific / General

< 1, C(1) > < Sunny Warm Normal Strong warm same , True / Yes / +ve / 1 >

Going from General to Specific,

Hypothesis G0(1) = C(1)

1 = 1

∴ G1 = G0

∴

S0

S1

G1

G0

Consider

<2, C(2)>

≡ < < sun

DATE: PAGE:

Going from Specific to General.

$S_0(1) \neq c(1)$

∴ we need to get a more general hypothesis S_1 such that

$S_1(1) = c(1)$

S_1 : < Sunny Warm Normal Strong Warm Same >

Now, $S_1(1) = c(1)$

EnjoySport	
Yes	(+ve)
yes	(+ve)
No	(-ve)
Yes	(+ve)

∴ For ex1,

S_0	< ϕ ϕ ϕ ϕ ϕ ϕ >
S_1	< S W N S W S >
G_1	< ? ? ? ? ? ? >
G_0	< ? ? ? ? ? ? >

Considering ex 2,

<2, c(2)>

≡ < < Sunny Warm High Strong Warm Same >

,

True / Yes / +ve / 1

>

DATE: PAGE:

Since it's a +ve example,

Start from most general to specific,

G1 < ? ? ? ? ? ? >

G1(ex1) = C(1)

∴ G2 = G1

G2 < ? ? ? ? ? ? >

From specific to general,

S1 < Sunny Warm Normal Strong Warm Same >

ex2 : < Sunny Warm High Strong Warm Same >

S1(2) ≠ C(2)

∴ Get a more general hypothesis such that it is consistent with both ex1 and ex2

⇓

And we know that S1 is already consistent with ex1

∴ Comparing S1 & ex2 attribute wise we get S2,

S2 : < Sunny Warm ? Strong Warm Same >

DATE: PAGE:

From ex1 & ex2 we get,

S_0 <ϕ ϕ ϕ ϕ ϕ ϕ>
S1 <S W N S W S>
S2 <S W ? S W S>

↓ General

G2 <? ? ? ? ? ?>
G1 <? ? ? ? ? ?>
G0 <? ? ? ? ? ?>

↑ Specific

Considering ex 3,

ex3 : Rainy cold High Strong Warm Change -ve

<3, c(3)>

< Rainy cold High Strong Warm Change,
False / No / -ve / 0 >

Since its a -ve example,
we'll start from Specific to General,

S2 and ex3 are compared attribute wise

∵ Sunny ≠ Rainy

∴ we get 0

and ∵ we have conjunctions

0 ∨ anything (0 or 1) ∨ anything ∨ anything ∨ anything ∨ anything

= 0

i.e., S2 designates ex3 as -ve example S2(3) = 0

and so is

c(3) = 0

∵ S2(3) = c(3) = 0

∴ S3 = S2

S3 <Sunny Warm ? Strong Warm same>

DATE: PAGE:

From general to specific,

G2 < ? ? ? ? ? ? >

ex 3 < Rainy cold High strong warm change >

c(3) = 0

G2(3) = 1

G2(3) ≠ c(3)

i.e., G2 is not consistent with ex 3

∴ we'll have to find a new more specific (than G2) hypothesis (G3) which is consistent with ex 1, 2 & 3.

i.e.,

G(1) = c(1) [= 1]

and

G(2) = c(2) [= 1]

and

G(3) = c(3) [= 0]

G3 : < Sunny ? ? ? ? ? >

So we get :

G3 : < Sunny warm ? Strong Warm Same >

G3 : { < Sunny ? ? ? ? ? >,
< ? Warm ? ? ? ? >,
< ? ? ? ? ? Same > }

Both S3 & G3 are consistent with ex 1, 2 & 3.

ex4 : Sunny Warm High Strong cold change DATE: +ve PAGE:

<4, c(4)>

<Sunny Warm High Strong cold change, 1>

Since its a +ve example,

we start from most general to specific.

Is ex4 consistent with G3 ?

G3

<Sunny ? ? ? ? ?> <? Warm ? ? ? ?> <? ? ? ? ? Same>

consistent consistent Not consistent

✓ ✓ ✗

∴ G4 = { <Sunny ? ? ? ? ?>

<? Warm ? ? ? ? > }

Is ex4 consistent with G3 ?

<Sunny Warm ? ? ? ?>

<Sunny ? ? Strong ? ? ?>

<? Warm ? Strong ? ? ?>

S3

<Sunny Warm ? Strong Warm Same>

No.

S4 = <Sunny Warm ? Strong ? ?>

Final Version Space, V :

S4 = <Sunny Warm ? Strong ? ?>

<Sunny Warm ? ? ? ? > <Sunny ? ? Strong ? ?> <? Warm ? Strong ? ?>

G4 = <Sunny ? ? ? ? ?> <? Warm ? ? ? ? >

Probability

DATE: PAGE:

1. A coin is tossed 4 times. P(exactly 1 head) = ?

Sol. $P(HTTT) + P(THTT) + P(TTHT) + P(TTTH)$

$= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) \times 4$

$= \frac{1}{4}$

2. A bag contains 6 Red & 4 Black Balls. 2 balls are drawn at random one after the other without replacement. P(2 black balls) = ?

Sol. $\frac{4}{10} \times \frac{3}{9} = \frac{2}{5} \times \frac{1}{3} = \frac{2}{15}$

3. A card is drawn randomly from deck. I win Rs 1000 if it is a diamond or an ace.
P(winning) = ?

Sol. $P(\text{winning}) = \frac{13+3}{52} = \frac{16}{52} = \frac{4}{13}$

4. Bag 1 contains 4 white & 6 black balls. Bag 2 contains 4 white & 3 black balls. P(picking black ball from bag 1)

Sol. $P = \frac{1}{2}\left(\frac{6}{10}\right) = \frac{3}{10}$

P(picking ball | bag 1) =

DATE: PAGE:

P (choosing bag 1 & black ball) = $\frac{1}{2}\left(\frac{6}{10}\right)$

P (" bag 2 & black ball) = $\frac{1}{2}\left(\frac{3}{7}\right)$

P (black ball) = P(choose bag 1) P(picking black ball | chose bag 1)

+

P(choose bag 2) P(Picking black ball | chose bag 2)

$= \frac{1}{2}\left(\frac{6}{10}\right) + \frac{1}{2}\left(\frac{3}{7}\right)$

$= \frac{1}{2}\left[\frac{6}{10} + \frac{3}{7}\right] = \frac{1}{2}\left[\frac{3}{5} + \frac{3}{7}\right]$

$= \frac{3}{2}\left[\frac{12}{35}\right] = \frac{3 \cdot 6}{35} = \frac{18}{35}$

P(picking black ball from bag 1)

$= \frac{\left(\frac{1}{2}\right)\left(\frac{6}{10}\right)}{P(\text{black ball})} = \left(\frac{3}{10}\right) \div \left(\frac{18}{35}\right)$

$= \frac{3}{10} \times \frac{35}{18} = \frac{3 \times 7}{2 \ 18} = \frac{1 \times 7}{2 \ 6} = \frac{7}{12}$

5. Tommorow is Ramesh's wedding in Rajasthan.
Weather forecast predicted that it rains 5 days a year.
It rained 5 days per year in the last decade.
Prediction is not perfect.
Whenever it rains he perfectly predicts 90% of the time.
" it doesn't rain he " " 10% " " "
What is the probability it'll rain tomorrow?

DATE: PAGE:

6- Scores in Math contest follows Normal Distribution.

Arvind scores 75 / 100.

Avg. score for the test is 60.

Std dev = 15.

Top 10% of the students are eligible for scholarship.

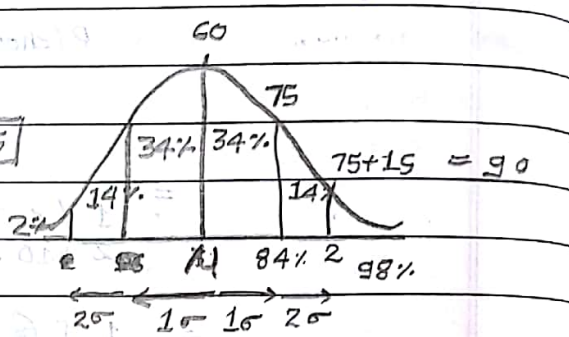
i) least score to qualify for scholarship.

ii) will Arvind get scholarship.?

iii) How many students are below Arvind ?

Sol.

i) $1.29 = x - 60 / 15 \Rightarrow \boxed{x = 79.35}$

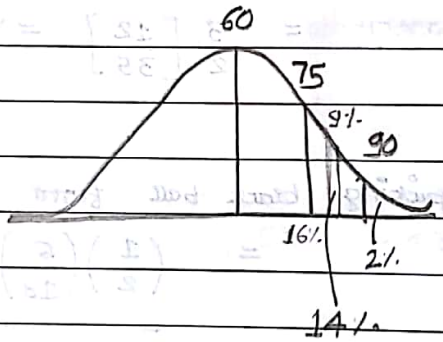


ii) No.

iii) 50 + 34

= 84 % of the students are below Arvind.

50 - 16 = 34 %.



75 + 9 = 84

or

90 - 9 = 81

1st September, 2017

Unit 2 DATE: PAGE:

Decision Trees

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	~~Overcast~~ Rain	Cool	Normal	Strong	No
D7	~~Sunny~~ Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	~~Rain~~ Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

NIR = MR = 5/14

- Decision Tree works with noisy training data as well.

$$E(S) = -\left[\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \frac{5}{14}\log_2\left(\frac{5}{14}\right)\right]$$

$$= -\left[\frac{9}{14}(-0.63743) + \frac{5}{14}(-1.48543)\right]$$

$$= -[-0.40977 - 0.53051]$$

$$= -[-0.9403]$$

$$= 0.94$$

$$\log_b a = \frac{\log_{10} a}{\log_{10} b}$$

$$\log_2 a = \frac{\log_{10} a}{\log_{10} 2}$$

DATE: PAGE:

Given the version space V for Enjoy Sport example:

< Sunny Warm ? Strong ? ? >

<Sunny ? ? Strong ? ?> <Sunny Warm ? ? ? ?> <? Warm ? Strong ? ? ?>

< Sunny ? ? ? ? ?> <? Warm ? ? ? ?>

Classify the Test Data:

Instance	Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy Sport
A	Sunny	Warm	Normal	Strong	Cool	Change	?
B	Rainy	Cold	Normal	Light	Warm	Same	?
C	Sunny	Warm	Normal	Light	Warm	Same	?
D	Sunny	Cold	Normal	Strong	Warm	Same	?

Sol - Instance A :

All 6 hypothesis give +ve .·. classification is +ve with 100% co

Instance B :

All 6 hypothesis give -ve . ∴ " " -ve " 100%

Instance C :

3 give -ve .

3 give +ve .

This is the case of ambiguity.

Our model (or learner) cannot classify this example with confidence.

NOTE : This example if were available as training data (i.e., if we knew the label of this instance) and then it would have been the most optimal training example

at the rate of $\log_2$ ||vs|| ⇓ ? ? > we can reduce the VS by that wi

Scanned by CamScanner

DATE: PAGE:

to reduce the current version space size & to come closer to the target fn.

at the rate of $\log_2 \|VS\|$? ?>

This is to be expected, because those instances whose classification is most ambiguous are precisely the instances whose true classification would provide the most new information for refining the version space.

we can reduce the VS by half with such a training example (Here it would have been from 6 hypothesis in VS to 3 " " " VS)

Instance D :

4 give −ve.

2 gives +ve.

st Enjoy Sport

ge ?

? Here we can classify it as −ve with $\frac{4}{6} \times 100$

? $= \frac{2}{3} \times 100$

? $= 66.66\%$

confidence

with 100% confidence

" 100% "

Method II :

Instance A :

compare it with the most specific hypothesis.

<sunny warm ? strong ? ?>

Since this classifies instance A as +ve we need not check with other hypothesis, as all other hypothesis are more general than this hypothesis. By dint of more general than (& by common sense) we can say that all other hypothesis will also classify it as +ve.

DATE: PAGE:

Instance B

Comparing it most specific hypothesis,
its classification is -ve.
We can't say about the true classification of instance B just because most specific hypothesis gave classification is -ve.

What we can do is
compare instance B with hypothesis in General Set.
If all hypothesis " " " classifies instance B as -ve,
we can say for sure that the true classification of instance B is -ve.

Why ?
Because every other hypothesis is more specific than this hypothesis. So if these say -ve, all others will surely classify it as -ve.

True classification of instance B is -ve.
with confidence 100 %.

Instance C

Most specific hypothesis classifies it as -ve
Hypothesis in general set " " " +ve & +ve.
Other three hypothesis " " " -ve, +ve & -ve

No classification can be inferred.

Instance D

Most specific hypothesis classifies it as -ve
Hypothesis in general set " " " +ve & -ve
Other three hypothesis " " " +ve, -ve & -ve.

True classification is -ve with $4 \times 100 = 66.66\%$ confidence.

Entropy characterizes the (im)purity of an arbitrary collection of examples

If S is the collection of samples then,

$$\text{Entropy}(S) = -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus}$$

where P_+ : Proportion of all +ve examples in S

P_- : " " " -ve " " S

In case,

1. All +ve examples $\Rightarrow P_{\oplus} = 1$

$P_{\ominus} = 0$

Then,

$$\begin{aligned}\text{Entropy}(S) &= -[P_{\oplus} \log P_{\oplus} + P_{\ominus} \log P_{\ominus}] \\ &= -[1 \log 1 + 0] \\ &= -[1(0) + 0] \\ &= -0 \\ &= 0\end{aligned}$$

2. All -ve examples $\Rightarrow P_{\oplus} = 0$

$P_{\ominus} = 1$

Then,

$$\begin{aligned}\text{Entropy}(S) &= -[0 + 1 \log 1] \\ &= -[0+0] \\ &= -0 \\ &= 0\end{aligned}$$

3. Equal # of +ve & -ve examples :

$P_{\oplus} = 0.5$, $P_{\ominus} = 0.5$

$$\begin{aligned}\text{Entropy}(S) &= -[P_{\oplus} \log_2 P_{\oplus} + P_{\ominus} \log_2 P_{\ominus}] \\ &= -\left[0.5 \log_2\left(\frac{1}{2}\right) + 0.5 \log_2\left(\frac{1}{2}\right)\right] \\ &= -2 \times 0.5 \log_2\left(\frac{1}{2}\right) \\ &= -1\left[\log_2 1 - \log_2 2\right] = +\log_2 2 = 1\end{aligned}$$

DATE: PAGE:

4. Unequal # of +ve & -ve | examples :

$0 < E(S) < 1$

Information Gain (Sample S, for attribute A) :

$$Gain(S, A) = Entropy(S) - \sum_{v \in values\ for\ A} \frac{|S_v|}{|S|} \times Entropy(S_v)$$

$$Gain(S, Outlook) = 0.94 - \left[\frac{|S_{sunny}|}{|S|} \times E(S_{sunny}) \right] + \frac{|S_{overcast}|}{|S|} \times E(S_{overcast}) + \left[\frac{|S_{rain}|}{|S|} \times E(S_{rain}) \right]$$

$$= 0.94 - \left[\frac{5}{14} \times -\left[\frac{3}{5} \log_2 \left(\frac{3}{5}\right) + \frac{2}{5} \log_2 \left(\frac{2}{5}\right) \right] \right] + \frac{4}{14} \times -\left[\frac{4}{4} \log_2 \left(\frac{4}{4}\right) + \frac{0}{4} \log_2 \left(\frac{0}{4}\right) \right] + \left[\frac{5}{14} \times -\left[\frac{3}{5} \log_2 \left(\frac{3}{5}\right) + \frac{2}{5} \log_2 \left(\frac{2}{5}\right) \right] \right]$$

$$= 0.94 - \left[\frac{-5}{14} (-0.44218 - 0.528771) + 0 + \frac{-5}{14} (-0.44218 - 0.528771) \right]$$

DATE: PAGE:

$= 0.94 - \left[\frac{-5}{14}(-0.97095) \times 2\right]$

$= 0.94 - [0.34677 \times 2]$

$= 0.94 - 0.6935365$

$= 0.24646$

Gain (S, outlook) $\simeq 0.247$

where,

$E(S_{sunny}) = -0.97$

$E(S_{overcast}) = 0$

$E(S_{rain}) = -0.97$

To compute Gain (S, temp) we need,

$E(S_{Hot}) = E[2+, 2-] = -\left[\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$

$= -\left[2 \times \frac{1}{2} \times \left(\log_2 1 - \log_2 2\right)\right]$

$= -[(0-1)]$

$E(Hot) = 1$

$E(S_{cool}) = E[3+, 1-] = -\left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$

$= -[-0.31128 - 0.5]$

$= 0.81128 \simeq 0.81$

$E(S_{Mild}) = E[4+, 2-] = -\left[\frac{4}{6}\log_2\left(\frac{4}{6}\right) + \left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right)\right]$

$= -[-0.3899 - 0.5283] = 0.9181$

$\simeq 0.92$

DATE: PAGE:

$$Gain(S, temp) = E(S) - \left[\frac{\|S_{hot}\|}{\|S\|} \times E(S_{hot}) + \frac{\|S_{cool}\|}{\|S\|} \times E(S_{cool}) + \frac{\|S_{mild}\|}{\|S\|} \times E(S_{mild}) \right]$$

$$= 0.94 - \left[\frac{4}{14}(1) + \frac{4}{14}(0.81) + \frac{6}{14}(0.92) \right]$$

$$= 0.94 - \left[0.2857 + \frac{2}{7}(0.81) + \frac{3}{7}(0.92) \right]$$

$$= 0.94 - [0.2857 + 0.2857(0.81) + 0.4285(0.92)]$$

$$= 0.94 - [0.2857 + 0.2314 + 0.39428]$$

$$= 0.94 - [0.9113844]$$

$$Gain(S, temp) \simeq 0.0286$$

To compute Gain (S, Humidity) :

$$E[S_{High}] = E[3+, 4-] = -\left[\frac{3}{7} \log_2\left(\frac{3}{7}\right) + \frac{4}{7} \log_2\left(\frac{4}{7}\right) \right]$$

$$= -\left[0.4857 \log_2\left(\frac{3}{7}\right) + 0.5714 \log_2\left(\frac{4}{7}\right) \right]$$

$$= -[-0.5237 - 0.46132] \simeq +0.985$$

DATE: PAGE:

$E[S_{Normal}] = E[6+, 1-]$

$= -\left[\frac{6}{7}\log\left(\frac{6}{7}\right) + \frac{1}{7}\log\left(\frac{1}{7}\right)\right]$

$= \frac{-}{\log 2}\left[-0.0574 + 0.142857\log\left(\frac{1}{7}\right)\right]$

$= \frac{-1}{\log 2}\left[-0.0574 - 0.120728\right]$

$= \frac{-1}{\log 2}\left[-0.178128\right]$

$= 0.5917$

$Gain(S, Humidity) = E(S) - \left[\frac{\|S_{Normal}\|}{\|S\|} \times E(S_{Normal}) + \frac{\|S_{High}\|}{\|S\|} \times E(S_{High})\right]$

$= E(S) - \left[\frac{7}{14}(0.5917) + \frac{7}{14}(-0.985)\right]$

$= 0.94 - \left[\frac{1}{2}(0.5917) + \frac{0.985}{2}\right]$

$= 0.94 - \left[(0.5)(0.5917) + (0.5)(0.985)\right]$

$= 0.94 - \left[0.29585 + 0.4925\right]$

$= 0.94 - \left[0.78835\right]$

$= 0.15165$

Scanned by CamScanner

DATE: PAGE:

To compute Gain (S, Wind) we need :

$$E[S_{weak}] = E[6+, 2-]$$

$$= -\left[\frac{6}{8}\log_2\left(\frac{6}{8}\right) + \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right]$$

$$= \frac{-1}{\log 2}\left[-0.0937 + -0.1505\right]$$

$$= \frac{-1}{\log 2}\left[-0.244215\right]$$

$$= \log 0.81126$$

$$E[S_{strong}] = E[3+, 3-] = -\left[\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right]$$

$$= -\left[\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$$

$$= -\left[\frac{2\times 1}{2}\left[\log_2(1) - \log_2 2\right]\right]$$

$$= -[0+1]$$

$$= +1$$

$$= 1$$

Information

Gain (S

Gain (S

Gain (

Gain

Maxim

DATE: PAGE:

Information Gain :

Gain (S , Outlook) = 0.247

Gain (S , Temp) = 0.0286

Gain (S , Humidity) = 0.15165

Gain (S , Wind) = $E(S) - \left[\frac{\|S_{weak}\|}{\|S\|} \times E(S_{weak}) + \frac{\|S_{strong}\|}{\|S\|} \times E(S_{strong}) \right]$

$= 0.94 - \left[\frac{8}{14} (0.81126) + \frac{6}{14} (1) \right]$

$= 0.94 - \left[\frac{4}{7} (0.81126) + \frac{6}{14} \right]$

$= 0.94 - \left[0.5714 (0.81126) + \frac{6}{14} \right]$

$= 0.94 - [0.46357 + 0.42857]$

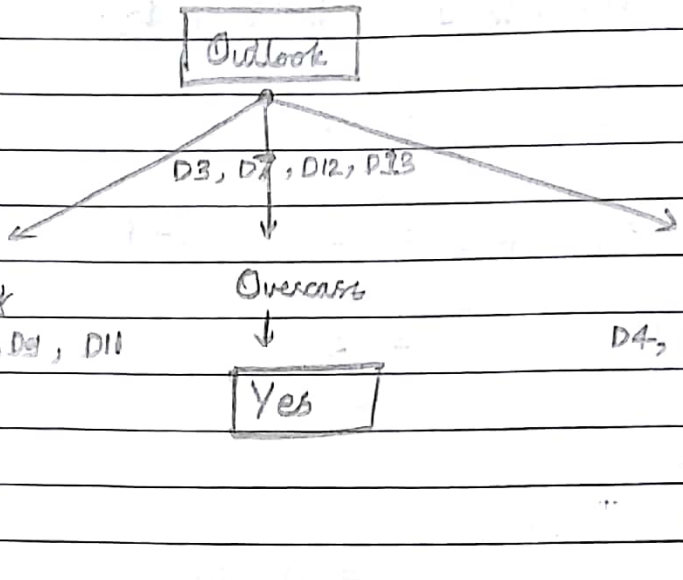
$= 0.94 - [0.89214]$

$= 0.0478$

Maximum Information Gain is with Outlook.

∴ Root Node will be Outlook.

DATE: PAGE:



Calculating for Sunny :

	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	weak	No
D2	Hot	High	Strong	No
D8	Mild	High	weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

$$E(\text{Sunny}) = -\left[\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]$$

$$= -\frac{1}{\log 2}\left[-0.15918 - 0.13311\right]$$

$$\simeq 0.971$$

Calculating Gain(Sunny, Temp) we need :

$$E(\text{Sunny}_{\text{Hot}}) = E[0+, 2-] = 0$$

DATE: PAGE:

$E[Sunny_{Mild}] = E[1+, 1-]$

$= 1$

$E[Sunny_{cool}] = E[1+, 0-]$

$= 0$

$Gain(Sunny, Temp) = E(Sunny) - \left[\frac{\|Sunny_{Hot}\|}{\|Sunny\|} \times E(Sunny)_{Hot}\right] + \frac{\|Sunny_{Mild}\|}{\|Sunny\|} \times E(Sunny_{Mild}) + \frac{\|Sunny_{cool}\|}{\|Sunny\|} \times E(Sunny_{cool})$

$= 0.971 - \left[\frac{2}{5}(0) + \frac{2}{5}(1) + \frac{1}{5}(0)\right]$

$= 0.971 - [0.40]$

$= 0.571$

Calculating $Gain(Sunny, Humidity)$ we need :

$E(Sunny_{High}) = -\left[0 + \frac{3}{3}\log_2\left(\frac{3}{3}\right)\right]$

$E[0+, 3-] = 0$

DATE: PAGE:

$E(\text{Sunny}_{\text{Normal}}) = E[2+, 0-]$

$= 0$

Gain (Sunny, Humidity)

$= E(\text{Sunny}) - \left[E(\text{Sunny}_{\text{High}}) \times \frac{\|\text{Sunny}_{\text{High}}\|}{\|\text{Sunny}\|} + E(\text{Sunny}_{\text{Normal}}) \times \frac{\|\text{Sunny}_{\text{Normal}}\|}{\|\text{Sunny}\|} \right]$

$= 0.971 - [0 + 0]$

$= 0.971$

Calculating Gain (Sunny, Wind) we need :

$E[\text{Sunny}_{\text{Strong}}] = E[1+, 1-]$

$= 1$

$E[\text{Sunny}_{\text{weak}}] = E[1+, 2-]$

$= -\left[\frac{1}{3} \log_2\left(\frac{1}{3}\right) + \frac{2}{3} \log_2\left(\frac{2}{3}\right) \right]$

$= -\frac{1}{\log 2} [-0.159 - 0.1174]$

$= 0.918$

Gain (Sunny, Wind) $= E(\text{Sunny}) - \left[\frac{\|\text{Sunny}_{\text{Strong}}\|}{\|\text{Sunny}\|} \times E[\text{Sunny}_{\text{Strong}}] + \frac{\|\text{Sunny}_{\text{weak}}\|}{\|\text{Sunny}\|} \times E[\text{Sunny}_{\text{weak}}] \right]$

$= 0.971 - \left[\frac{2}{5}(1) + \frac{3}{5}(0.918) \right] = 0.971 - (0.4 + 0.8262) = 0.4882$

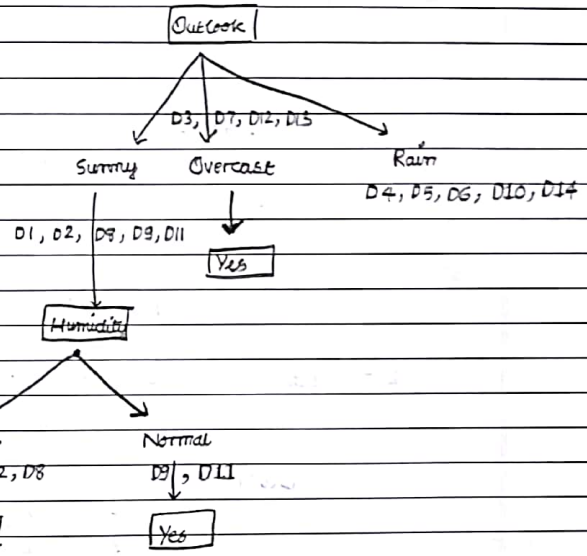
DATE: PAGE:

Gain (Sunny, Humidity) = 0.971

Gain (Sunny, Wind) = 0.138

Maximum Information Gain is with Humidity

∴ Next Node in the tree is Humidity



DATE: PAGE:

Calculating for Rainy :

	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

$$E(\text{Rainy}) = E[3+, 2-] = -\left[\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

$$= \frac{-1}{\log 2}\left[0.6\,(-0.22185) + 0.4\,(-0.3979)\right]$$

$$= \frac{-1}{\log 2}\left[-0.13311 - 0.15916\right]$$

$$= 0.97089$$

$$\simeq 0.979$$

Calculating Gain (Rainy, Temp) we need :

$$E(\text{Rainy}_{\text{Mild}}) = E[2+, 1-]$$

$$= -\left[\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$$

DATE: PAGE:

$$= \frac{-1}{\log 2}\left[-0.116 - 0.1574\right]$$

$$= 0.9084$$

$$E(\text{Rainy}_{\text{Cool}}) = E[1+, 1-]$$

$$= 1$$

$$\text{Gain}(\text{Rainy}, \text{Temp}) = E(\text{Rainy}) - \left[\frac{3}{5}(0.9084) + \frac{2}{5}(1)\right]$$

$$= 0.979 - \left[0.54504 + 0.4\right]$$

$$= 0.979 - 0.94504$$

$$= 0.03396$$

916]

Calculating Gain (Rainy, Humidity) we need :

$$E(\text{Rainy}_{\text{High}}) = E[1+, 1-]$$

$$= 1$$

$$E(\text{Rainy}_{\text{Normal}}) = E[2+, 1-]$$

$$= 0.9084$$

$$\text{Gain}(\text{Rainy}, \text{Humidity}) = 0.979 - \left[\frac{2}{5}(1) + \frac{3}{5}(0.9084)\right]$$

$$= 0.979 - [0.4 + 0.54504] = 0.03396$$

DATE: PAGE:

Calculating Gain (Rainy, Wind) we need :

$E(Rainy_{Weak}) = E[3+, 0-]$

$= 0$

$E(Rainy_{Strong}) = E[0+, 2-]$

$= 0$

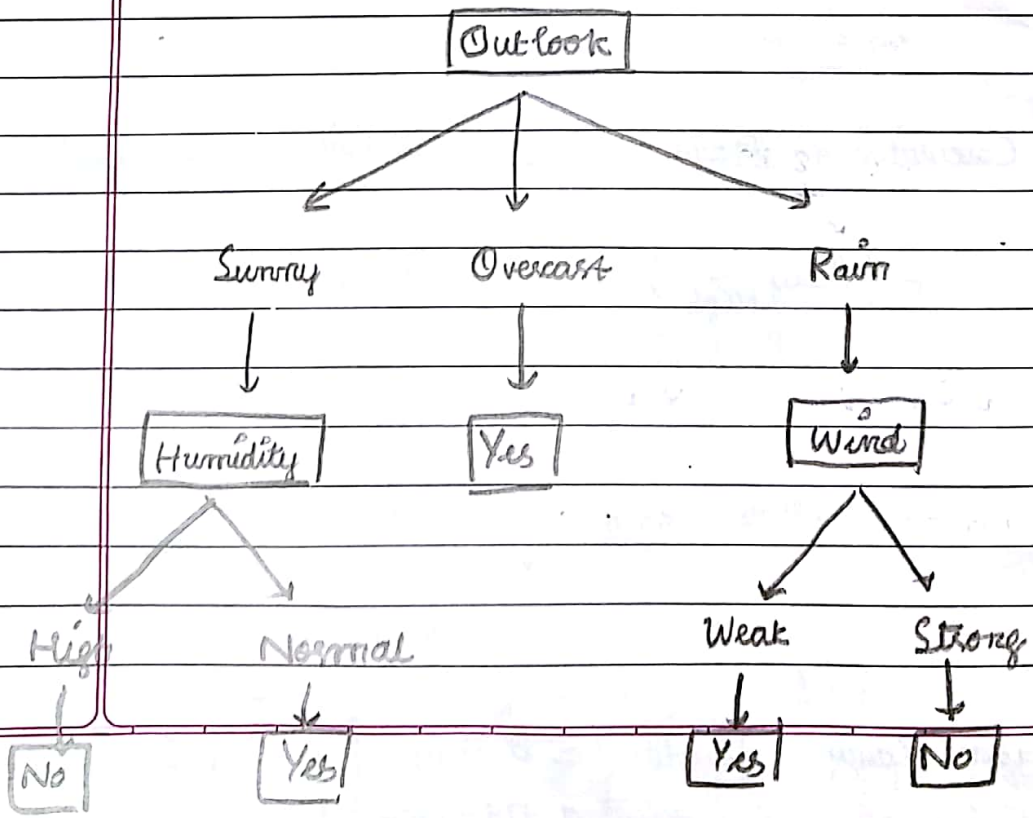
Gain (Rain, Wind) = E(Rain) − [0 + 0]

= 0.979 − 0

= 0.979

~~Information Gain is max with wind~~

∴ Next node in the tree will be wind



Decision Trees Home-work : DATE: PAGE:

Instance	Gender	Age	Country	Recommendation
D1	Male	Young	USA	Action (+)
D2	Female	Young	India	Action (+)
D3	Male	Old	USA	Drama (-)
D4	Female	Old	India	Action (+)
D5	Female	Young	USA	Action (+)
D6	Male	Old	India	Drama (-)
D7	Female	Old	China	Drama (-)
D8	Male	Young	China	Action (+)

Entropy, $E(S) = -\left[p_{Action} \log_2 p_{Action} + p_{Drama} \log_2 p_{Drama}\right]$

$= -\left[\left(\frac{5}{8}\right)\log_2\left(\frac{5}{8}\right) + \left(\frac{3}{8}\right)\log_2\left(\frac{3}{8}\right)\right]$

$= -\left[0.625 \log_2\left(\frac{5}{8}\right) + (0.375)\log_2\left(\frac{3}{8}\right)\right]$

$= -\frac{1}{\log_{10} 2}\left[-0.1275 - 0.15974\right]$

$= \frac{-0.28724}{\log_{10} 2} = 0.9542$

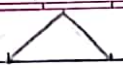
Information Gain (Sample S, For attribute A):

$$Gain(S, A) = E(S) - \sum_{v \in \text{values for } A} \frac{\|S_v\|}{\|S\|} \times E(S_v)$$

For attribute Gender, Sample S

DATE: PAGE:

Given sa



Male (4) [2+, 2-]

Female (4) [3+, 1-]

$E(S, Gender_{Male}) = 1$

$E(S, Gender_{Female}) = -\left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$

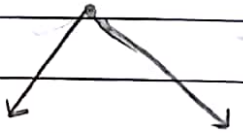
$= -\frac{1}{\log_{10} 2}\left[0.75(-0.12494) + 0.25(-0.60205)\right]$

$= -\frac{1}{\log_{10} 2}\left[-0.093705 - 0.150525\right]$

$= 0.81132$

Info

Given sample S, for attribute Age



Young (4) [4+, 0-]

Old (4) [1+, 3-]

$E(S, Age_{Young}) = 0$

$E(S, Age_{old}) = 0.81132$

Given sample S, for attribute Country DATE: PAGE:

USA(3) India(3) China(2)
[2+, 1-] [2+, 1-] [1+, 1-]

$E(S, \text{Country}_{USA}) = -\left[\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$

$E(S, \text{Country}_{India}) = 0.9091$

$E(S, \text{Country}_{China}) = 1$

$= -\frac{1}{\log_{1.0} 2}\left[0.66 \log_{1.0}\left(\frac{2}{3}\right) + 0.33 \log_{1.0}\left(\frac{1}{3}\right)\right]$

$= -(3.3219)\left[-0.11622 - 0.15745\right]$

$= -(3.3219)(0.27367)$

$= 0.9091$

Information Gain (Sample S, for attribute A) :

Info Gain (S, Gender) $= E(S) - \left[\frac{\|S_{Male}\|}{\|S\|} \times E(S, \text{Gender}_{Male}) + \frac{\|S_{Female}\|}{\|S\|} \times E(S, \text{Gender}_{Female})\right]$

$= 0.9542 - \left[\frac{4}{8}(1) + \frac{4}{8}(0.81132)\right]$

$= 0.9542 - \left[0.5 + 0.40516\right]$

$= 0.9542 - 0.90516 = 0.04904$

Scanned by CamScanner

DATE: PAGE:

- Info Gain (S, Age) = $0.9542 - \left[\frac{4}{8}(0) + \frac{4}{8}(0.81128)\right]$

$= 0.9542 - 0.40566$

$= 0.54854$

- Info Gain (S, Country) = $0.9542 - \left[\frac{3}{8}(0.9091) + \frac{3}{8}(0.9091) + \frac{2}{8}(1)\right]$

$= 0.9542 - \left[0.375(0.9091) + 0.375(0.9091) + 0.25\right]$

$= 0.9542 - [0.681825 + 0.25]$

$= 0.9542 - [0.931825]$

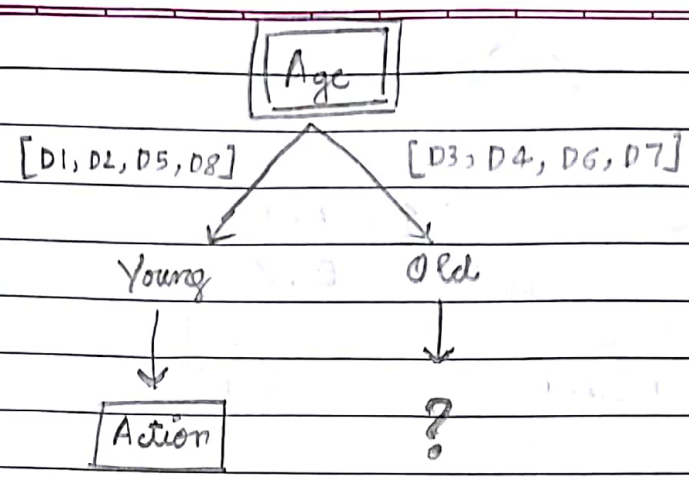
$= 0.022375$

Info Gain is maximum for Age

∴ Age becomes the root node.

Samp

DATE: PAGE:



Given Old, the Sample S becomes :

Sample T

Instance	Gender	Country	Recommendation
D3	Male	USA	Drama (-)
D4	Female	India	Action (+)
D6	Male	India	Drama (-)
D7	Female	China	Drama (-)

$$E(T) = -\left[p_{Action} \log_2 p_{Action} + p_{Drama} \log_2 p_{Drama} \right]$$

$$= -\left[\frac{1}{4} \log_2 \left(\frac{1}{4}\right) + \frac{3}{4} \log_2 \left(\frac{3}{4}\right) \right]$$

$$= \frac{-1}{\log_{10} 2} \left[0.25 \left(0 - 2\right) + 0.75 \left(\frac{\log 3}{2} - 2 \right) \right]$$

$$= -3.3219 \left(-0.5 + 0.75 \left(1.58496 - 2 \right) \right)$$

$$= 0.81128$$

DATE: PAGE:

For sample P, Gender (4)

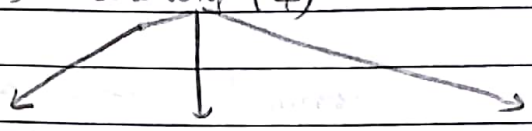
Male (2) [0+, 2-]

Female (2) [1+, 1-]

$E(P, Gender_{Male}) = 0$

$E(P, Gender_{Female}) = 1$

For sample P, Country (4)



USA (1) [0+, 1-]

India (2) [1+, 1-]

China (1) [0+, 1-]

$E(P, Country_{USA}) = 0$

$E(P, Country_{India}) = 1$

$E(P, Country_{China}) = 0$

Information Gain (P, Gender)

$$= 0.81128 - \left[\frac{2}{4}(0) + \frac{2}{4}(1)\right]$$

$$= 0.81128 - (0 + 0.5)$$

$$= 0.81128 - 0.5$$

$$= 0.31128$$

DATE: PAGE:

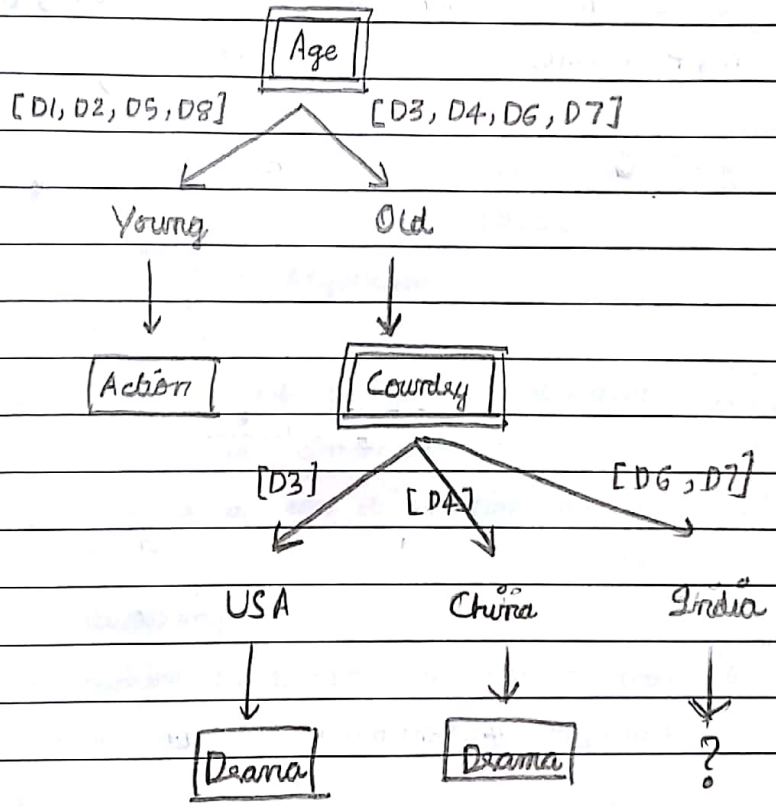
Info Gain (P, Country) = 0.81128 − [1(0)+1(0) / 4 4 + 2/4 (1)]

= 0.81128 − 0.5

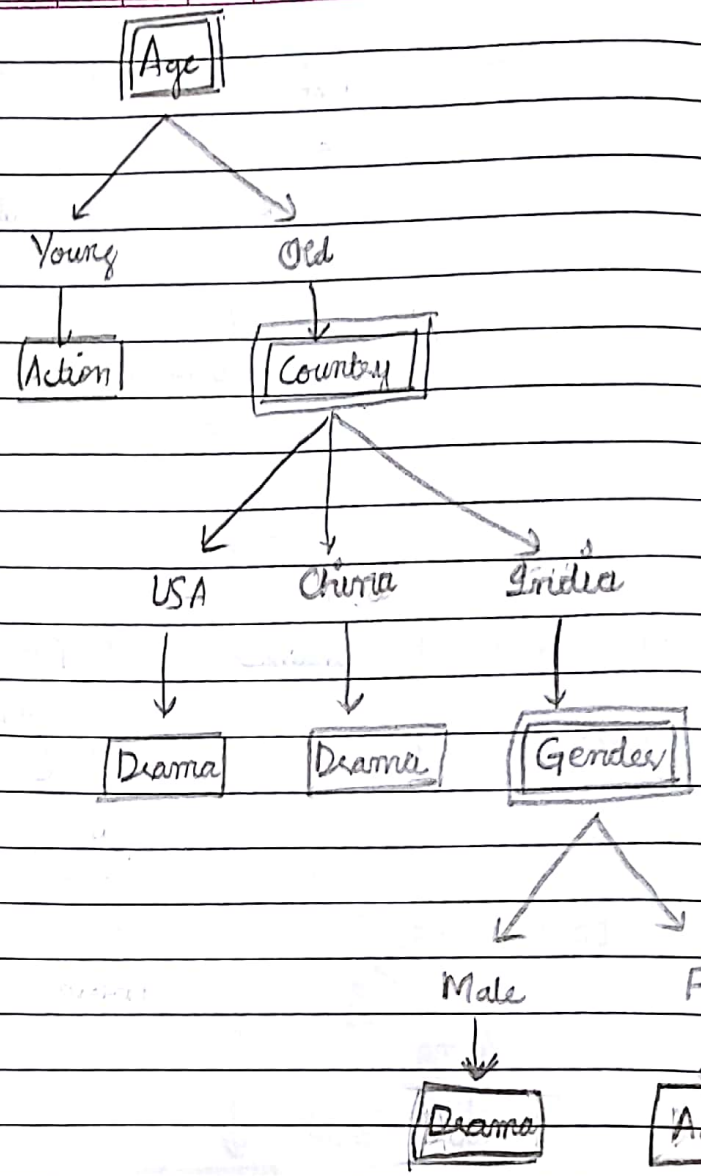
= 0.31128

Info Gain (P, Country) = Info Gain (P, Gender)

Choosing Country arbitrarily, we get



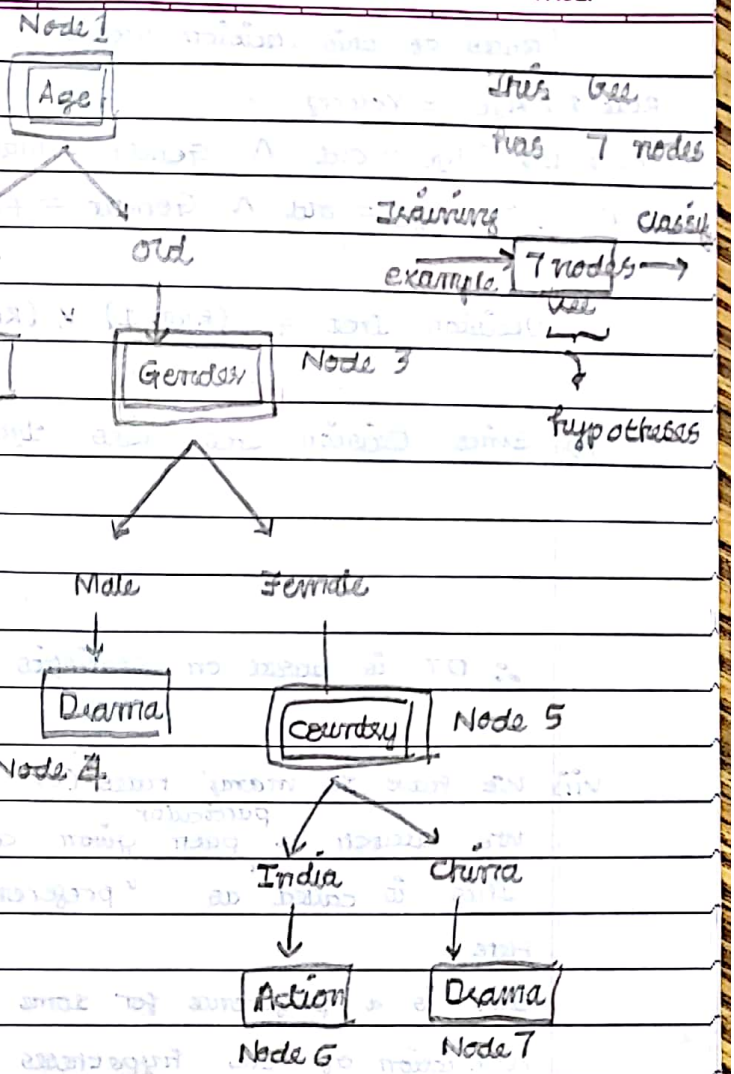
DATE: PAGE:



Old Indian women like action movies ..:P

Tree can also be :

DATE: PAGE:



This tree has 7 nodes

Training example → 7 nodes → Tree → hypotheses

Class

Decision Tree : ID3 Algorithm

i) Path from root to leaf : Rule or antecedant. Each rule is conjunction

ii) Leaf is called as consequent.

iii) Decision Tree is a disjunction of conjunctions. (path or rule)

iv) Bias : Searching

v) Once a decision to proceed towards a path is taken you can't backtrack — Greedy Approach

i.e.,

if your Test Instance is

Old Male India

⇓

once you have gone down the tree after looking at Old we can't go back. We will go to Gender now.

DATE: PAGE:

Rules of this decision tree :

Rule 1 : Age = Young

Rule 2 : (Age = old ∧ Gender = Male)

Rule 3 : (Age = old ∧ Gender = Female ∧ Country = India)

Decision Tree = (Rule 1) ∨ (Rule 2) ∨ (Rule 3)

vi) Since Decision Tree uses Information Gain

⇓

Statistical entity

∴ DT is based on statistics.

vii) We have so many rules (or paths).

We search a particular path given an instance.

This is called as "preference bias".

Here,

Bias is a preference for some hypotheses, rather than a restriction of the hypothesis space (Candidate Elimination Algorithm)

viii) If Training 40 examples

500 nodes Tree → classify

Occam's Razor

hypotheses h'

If for same 40 training examples,

if we get a (can) 20 nodes tree is obtained

hypotheses h

then we prefer h over h'. This is the bias we have on preferring h over h' based on Information Gain.

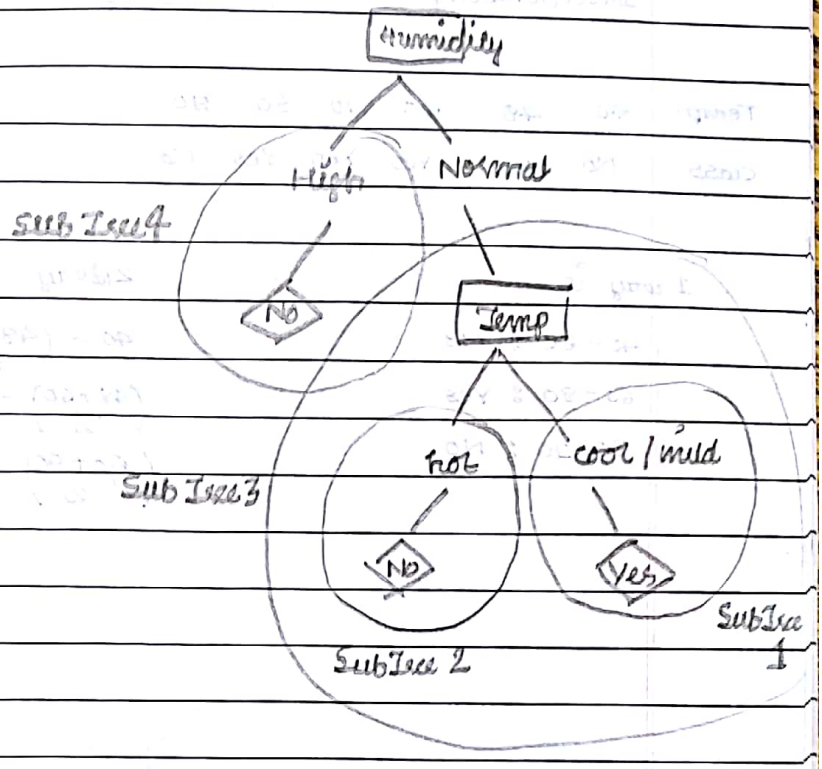
DATE: PAGE:

2 methods to improve accuracy :

1- Reduced Error Pruning : Removing subtrees to improve accuracy.

2- Rule based " : " rule (or path or antecedant)

Reduced Error Pruning :



And so on we have multiple subtrees.

We ~~~~ choose subtree 1 & remove it.

Check with all the training examples again.

See if this pruned tree classifies well.

Similarly, we iteratively try to prune all subtrees one by one & try with training examples.

See if after pruning subtree if accuracy is better.

This way we ~~~ get better accuracy

than we keep the shortest tree.

DATE: PAGE:

Rule based Pruning :

We prune rule one by one (rule or path).

If accuracy is better, remove the rule (or path).

Else retain.

Incorporating Continuous Values :

Temp.	40	48	60	70	80	90
Class	No	No	Yes	Yes	Yes	No

1 way is

40 - 60 : No

60 - 80 : Yes

80 - 90 : No

2 way

$40 - \left(\frac{48+60}{2}\right)$: No

$\left(\frac{48+60}{2}\right) - \left(\frac{80+90}{2}\right)$: Yes

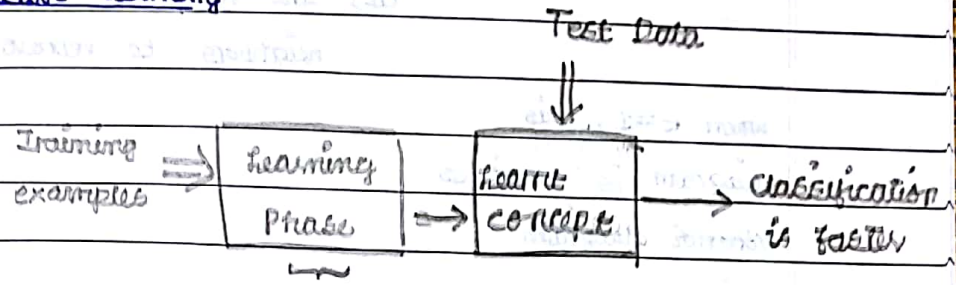
$\left(\frac{80+90}{2}\right) - 90$: No

(Both are Supervised Learning Algo)

DATE: PAGE:

Inductive Learning vs Instance Based Learning

Inductive Learning



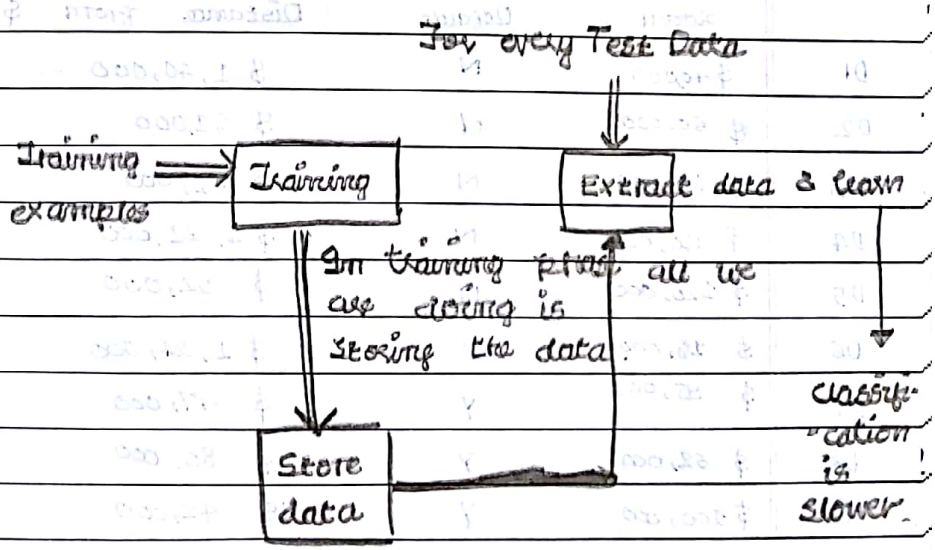
Consumes / takes time

Example algorithms are :

i) Concept Learning / Inductive Learning using Candidate Elimination Algorithm.

ii) " " / " " " Decision Tree.

Instance Based Learning : (Lazy Learning)



Example algorithm :

KNN or K - nearest neighbors

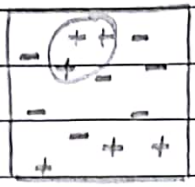
DATE: PAGE:

k nearest neighbors :

Requires 3 things : i) The set of stored records.

ii) Distance Metric to compute distance b/w records

iii) The value of k, the # of nearest neighbors to retrieve.



When k=1, this diagram is called as Voronoi diagram

For classification/ regression when, k ≥ 1 :

i) In classification, take the most frequent classification.

ii) In regression, find the average of all the values.

Example :

	Loan	Default	Distance from $ 142,000
D1	$40,000	N	$ 1,40,000
D2	$ 60,000	N	$ 82,000
D3	$ 80,000	N	$ 62,000
D4	$ 20,000	N	$ 1, 22,000
D5	$ 120,000	N	$ 22,000
D6	$ 18,000	Y	$ 1, 24,000
D7	$ 95,000	Y	$ 47,000
D8	$ 62,000	Y	$ 80,000
D9	$ 100,000	Y	$ 42,000
D10	$ 220,000	Y	$ 78,000
D11	$ 150,000	Y	$ 8000

If k = 5,

8K 47K 42K

DATE: PAGE:

D11, D7 & D9 are Y

D5 & D3 are N

∴ Classified as Y

22K 62K

rest

Ex 2 :

#	Action Scene	Romantic Scene	Emotional Scene	Classify (Movie)
D1	5	25	5	Romantic
D2	10	5	20	Emotional
D3	15	2	20	Emotional
D4	15	20	18	Romantic
D5	25	5	20	Action
D6	30	15	5	Action
Test Case	15	15	10	?

Distances of Test Case from Stored Instances :

D1 $\sqrt{(15-5)^2 + (15-25)^2 + (10-5)^2} = \sqrt{(10)^2+(10)^2+(5)^2} = 15$

D2 $\sqrt{(15-10)^2 + (15-5)^2 + (10-20)^2} = \sqrt{5^2+10^2+10^2} = 15$

D3 $\sqrt{(15-15)^2 + (15-2)^2 + (10-20)^2} = \sqrt{0+13^2+10^2} = 16.4$

D4 $\sqrt{(15-15)^2 + (15-20)^2 + (10-18)^2} = \sqrt{0+5^2+8^2} = 9.43$

D5 $\sqrt{(15-25)^2 + (15-5)^2 + (10-20)^2} = \sqrt{10^2+10^2+10^2} = 17.32$

D6 $\sqrt{(15-30)^2 + (15-15)^2 + (10-5)^2} = \sqrt{15^2+0+5^2} = 15.811$

Rom Emo Rom

For $k=3$, we get D1, D2 & D4 ∴ Test case is classified as Romantic

DATE: PAGE:

Algorithmically,

after we got D1, D2 & D4 (∵ we decided $k = 3$)

how do we classify the test case?

δ (A

For attribute Action Scene (A),

D1	classifies	Romantic Movie	$\delta(A, RM) +$
D2	"	Emotional "	$\delta(A, EM) +$
D4	"	Romantic "	$\delta(A, RM)$
			$= 0 + 0 + 0 = 0$

For attribute Romantic Scene (R),

D1	classifies	Romantic Movie	$\delta(R, RM) +$
D2	"	Emotional "	$\delta(R, EM) +$
D4	"	Romantic "	$\delta(R, RM)$
			$= 1 + 0 + 1 = 2$

Similarly for attribute Emotional Scene (E),

$\delta(E, RM) + \delta(E, EM) + \delta(E, RM)$

$=$

$0 + 1 + 0$

$=$

1

$\hat{f}(x_{query})$ $\max\{0, 2, 1\} = 2$ ∴ Romantic Movie is the classification.

$$\hat{f}(x_{query}) = \hat{f}(x_{test}) = \underset{v \in values}{\operatorname{argmax}} \sum_{i=1}^{k} \delta(v, f(x_i))$$

DATE: PAGE:

Better measures to calculate class from k neighbours :

$$w_i = \frac{1}{d(x_i, x_{query})}$$

$$\hat{f}(x_{query}) = \underset{v \in values}{\operatorname{argmax}} \sum_{i=1}^{k} w_i \, \delta(v, f(x_i))$$

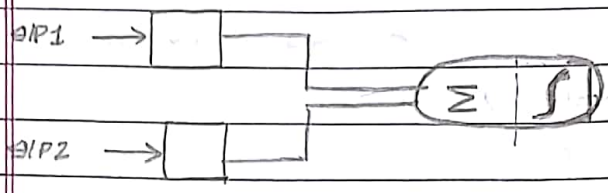
Exam question

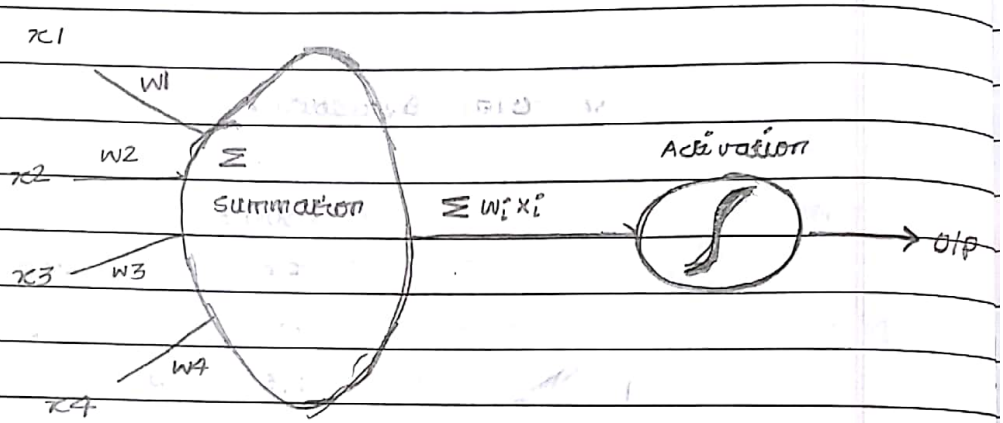
Artificial Neural Network (Concept learning)

DATE:

PAGE:

Perceptron : Artificial Neural

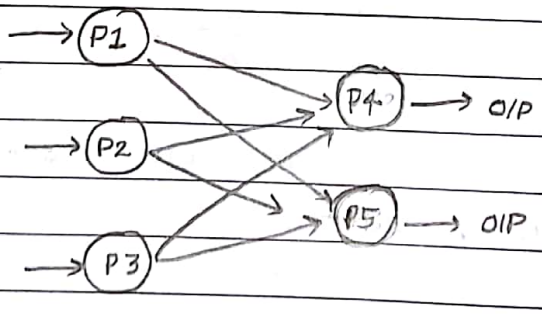




Activation Fn $\begin{cases} = 1, & \text{if } w_i x_i > 0 \\ = -1, & \text{if } w_i x_i \leq 0 \end{cases}$

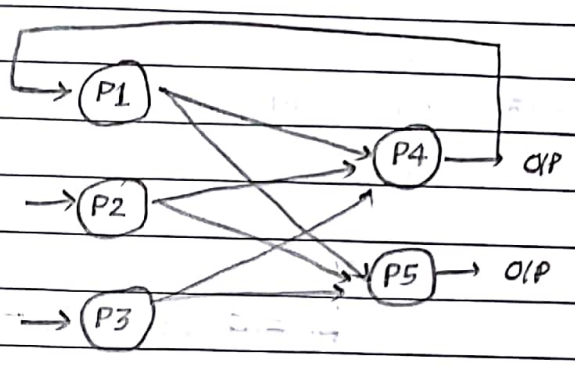
ANN :

Type 1 : Forward Neural Network



DATE: PAGE:

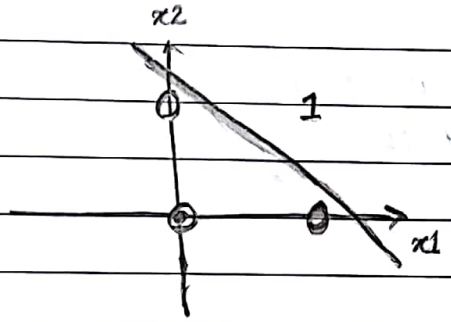
Type 2 : Feedback Forward Network



Linear Separation using Perceptron :

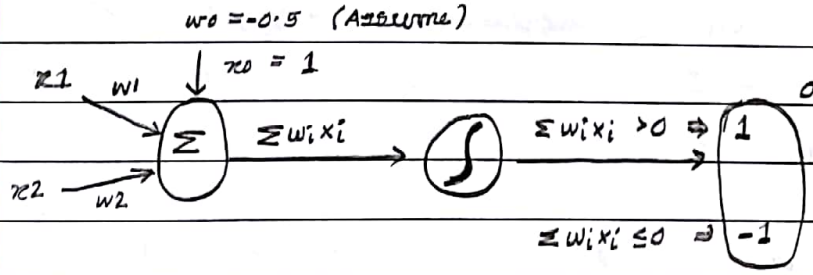
1. AND Gate :

Instance	x1	x2	O/P
d1	0	0	0
d2	0	1	0
d3	1	0	0
d4	1	1	1



We see that this data is linearly separable.

∴ Perceptron can be used to implement AND Gate.



Initializing w_1 & w_2 with random values we get :

Let, $w1 = 1$

$w2 = 0.5$

DATE: PAGE:

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

For instance 1

$\sum w_i x_i = -0.5(1) + 1(0) + (0.5)(0)$

$= -0.5$

∴ Activation $(-0.5) = -1$

When -1, O/P $= 0$

$\Rightarrow \hat{y}_1 = 0$

For instance d1,

True label, $y_1 = 0$ [0 && 0 = 0]

Predicted label, $\hat{y}_1 = 0$

Error = 0 ∴ No need to update the weights

For instance 2 : (0, 1)

$\sum w_i x_i = w_0 x_0 + w_1 x_1 + w_2 x_2$

$= (-0.5)(1) + (1)(0) + (0.5)(1)$

$= -0.5 + 0.5$

$= 0$

Activation $(0) = -1$

Updating

When O/P of activation is -ve,

O/P = 0

$\Rightarrow \hat{y}_2 = 0$

For instance d2,

Updating

True label, $y_2 = 0$ [0 && 1 = 0]

Predicted label, $\hat{y}_2 = 0$

Error = 0

∴ No need to update the weights.

Updatin

DATE: PAGE:

For instance 3, (1, 0)

$\Sigma w_i x_i = (-0.5)(1) + (1)(1) + (0.5)(0)$

$= -0.5 + 1$

$= 0.5$

Activation $(\Sigma w_i x_i) = 1$

If Activation $(\Sigma w_i x_i) \geq 0$, (+ve) O/P = 1

$\Rightarrow \hat{y}_3 = 1$

For instance d3,

true label, $y_3 = 0$ [1 & 0 = 0]

predicted label, $\hat{y_3} = 1$

Error = -1

∴ Perceptron did not predict correctly

∴ Update the weights.

To update the weights for instances d3 $(x_0, x_1, x_2) = (1, 1, 0)$,

$w_i = w_i + (\text{true label} - \text{predicted label}) x_i$

Assuming $\eta = 0.2$

Updating w0, $w_0 = w_0 + \eta(y_3 - \hat{y_3}) x_0$

$= (-0.5) + 2(0 - 1)(1)$

$= (-0.5) - (0.2)$

$= -0.7$

Updating w1, $w_1 = w_1 + \eta(y_3 - \hat{y_3}) x_1$

$= 1 + 0.2(0 - 1)(1)$

$= 1 - 0.2$

$= 0.8$

Updating w2, $w_2 = w_2 + \eta(y_3 - \hat{y_3}) x_2$

$= 0.5 + (0.2)(0 - 1)(0)$

$= 0.5$

DATE: PAGE:

Checking for instance d1, d2 and d3,

For instance 1 (1, 0, 0) with updated weights :

$\Sigma w_i x_i = (-0.7)(1) + (0.8)(0) + (0.5)(0)$

$= -0.7$

Activation $(\Sigma w_i x_i)$ = Activation (-0.7)

$= -1$

When o/p of activation is -ve,

predicted label $\hat{y}1 = 0$

True label $y1 = 0$

Error = 0 :)

For instance 2 (1, 0, 1) with updated weights :

$\Sigma w_i x_i = (-0.7)(1) + (0.8)(0) + (0.5)(1)$

$= -0.7 + 0.5$

$= -0.2$

Activation $(\Sigma w_i x_i) = -1$:

When o/p of activation is -ve, predicted label = 0

True label = 0

Error = 0 :)

For instance 3 (1, 1, 0) with updated weights :

$\Sigma w_i x_i = (-0.7)(1) + (0.8)(1) + (0.5)(0)$

$= -0.7 + 0.8 + 0$

$= 0.1$

Activation $(0.1) = 1$

DATE: PAGE:

With o/p of activation is +ve, predicted label $\hat{y_3} = 1$

true label $y_3 = 0$

[0 ≠ 0 = 0]

Error = −1 :(

∴ we need to update the weights with instance 3 one more time.

Updating weights with instance d3 (1, 1, 0) :

$w_0 = w_0 + \eta (y_3 - \hat{y_3}) x_0$

$= (-0.7) + 0.2 (0 - 1)(1)$

$= (-0.7) - 0.2$

$= -0.9$

$w_1 = w_1 + \eta (y_3 - \hat{y_3}) x_1$

$= (0.8) + 0.2 (0 - 1)(1)$

$= (0.8) - 0.2 = 0.6$

$w_2 = w_2 + \eta (y_3 - \hat{y_3}) x_2$

$= 0.5 + 0.2(0 - 1)(0)$

$= 0.5 - 0.$

$= 0.5$

0

0 [0 ≠ 1 = 0]

with $(w_0, w_1, w_2) = (-0.9, 0.6, 0.5)$

Instance (d_k)	$w_0 = -0.9$	$w_1 = 0.6$	$w_2 = 0.5$	Σ	Act(Σ)	$\hat{y}_k$
d1 (1,0,0)	-0.9	0	0	-0.9	-1	0
d2 (1,0,1)	-0.9	0	0.5	-0.4	-1	0
d3 (1,1,0)	-0.9	0.6	0	-0.3	-1	0

For d1, d2, d3

predicted label = true label

DATE: PAGE:

Using new $(w_0, w_1, w_2) = (-0.9, 0.6, 0.5)$

for instance d_4 $(1, 1, 1)$

Instance (d_4)	w_0	w_1	w_2	Σ	Activation(Σ)	$\hat{y}_4$	y_4	Error
$d_4(1, 1, 1)$	-0.9	0.6	0.5	$-0.9+1.1 = 0.2$	1	1	1	0 :)

∴ Final learnt weights,

$w_0 = -0.9$

$w_1 = 0.6$

$w_2 = 0.5$

Approximate target fn, $\hat{y} = -0.9 + 0.6x_1 + 0.5x_2$

for AND gate with initial $w_0 = 0.5$

$w_1 = 1$

$w_2 = 0.5$

and $\eta = 0.5$

weights were updated twice.

Ex 2 :

x_1	x_2	true label
1	1	1 (*)
2	-2	0 (△)
-1	-1.5	0 (△)
-2	-1	0 (△)
-2	1	1 (*)
1.5	-0.5	1 (*)

Assume initial values as,

$w_0 = 0$, $x_0 = 1$

$w_1 = 1$,

$w_2 = 0.5$

$\eta = 0.2$

Use $w_i = w_i + \eta$ (true label − predicted label) x_i

DATE: PAGE:

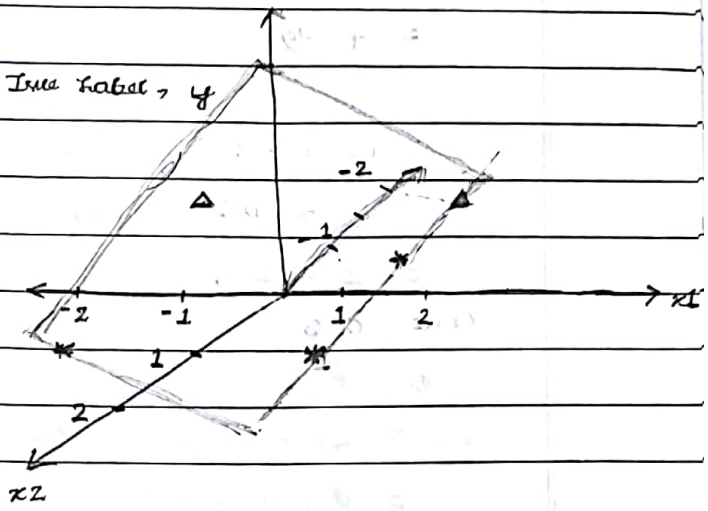
using a single perceptron.

Sol: A single perceptron can only approximate target fn for data which is linearly separable.

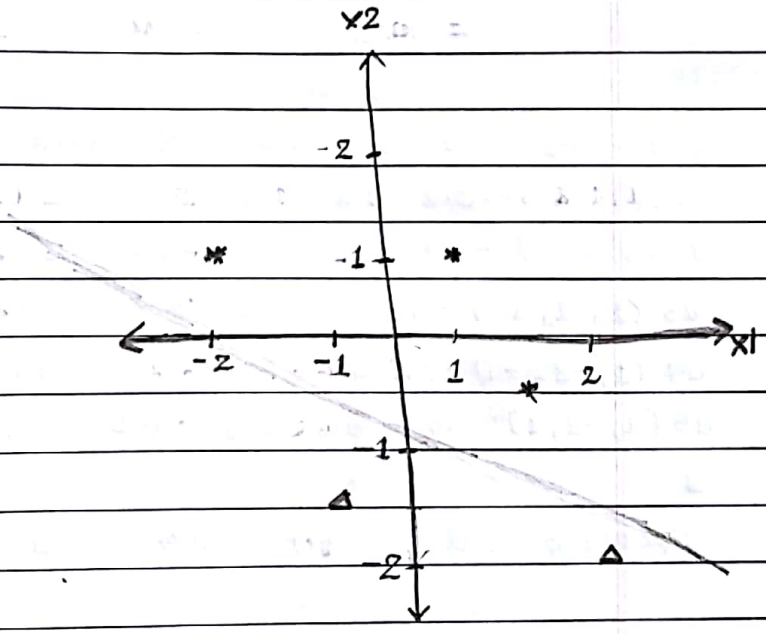
Error 0 :)

Checking if the given data is linearly separable,

Representation 1 ;



Representation 2 :



Yes, it is linearly separable.

∴ A single perceptron can be used to approximate the target fn.

∴ Target fn is of the form, $\hat{y} = w_0 + w_1x_1 + w_2x_2$

DATE: PAGE:

Instance d_k	w_0	w_1	w_2	Σ	Activation (Σ)	Predicted label, $\hat{y}_k$	True label, y_k	Cost $(y_k - \hat{y}_k)$
d1 (1,1,1)	0	1	0.5	1.5	1 ($\Sigma>0$)	1	1	0 :)
d2 (1,2,-2)	0	1	0.5	1	1 ($\Sigma>0$)	1	0	-1 :(

Updation of weights required with instance d2 (1, 2, -2)

$w_0 = w_0 + \eta(y_2 - \hat{y}_2)x_0$
$= 0 + 0.2(-1)(1)$
$= -0.2$

$w_1 = w_1 + \eta(y_2 - \hat{y}_2)x_1$
$= 1 + 0.2(-1)(2)$
$= 1 - 0.4$
$= 0.6$

$w_2 = w_2 + \eta(y_2 - \hat{y}_2)x_2$
$= 0.5 + 0.2(-1)(-2)$
$= 0.5 + 0.4$
$= 0.9$

Instance d_k	w_0	w_1	w_2	Σ	Activation (Σ)	$\hat{y}_k$	y_k	Error/Cost
d1 (1,1,1)	-0.2	0.6	0.9	1.3	1 ($\Sigma>0$)	1	1	0 :)
d2 (1,2,-2)	-0.2	0.6	0.9	-0.8	0 ($\Sigma\leq 0$)	0	0	0 :)
d3 (1,-1,-1.5)	-0.2	0.6	0.9	-2.15	0 ($\Sigma\leq 0$)	0	0	0 :)
d4 (1,-2,-1)	-0.2	0.6	0.9	-2.3	0 ($\Sigma\leq 0$)	0	0	0 :)
d5 (1,-2,1)	-0.2	0.6	0.9	-0.5	0 ($\Sigma\leq 0$)	0	1	-1 :(

Updating weights for instance d5 (1, -2, 1) :

$w_0 = w_0 + \eta(y_5 - \hat{y}_5)x_0$
$= -0.2 + 0.2(1)(1)$
$= -0.2 + 0.2$
$= 0$

$w_1 = w_1 + \eta(y_5 - \hat{y}_5)x_1$
$= 0.6 + 0.2(1)(-2) = 0.6 - 0.4 = 0.2$

$w_2 = w_2 + \eta(y_5 - \hat{y}_5)x_2$
$= 0.9 + 0.2(1)(1)$
$= 0.9 + 0.2$
$= 1.1$

Instance d_k	w_0	w_1	w_2	Σ	Activation (Σ)	$\hat{y}_k$	y_k	Error/Cost
d1 (1,1,1)	0	0.2	1.1	1.3	1 ($\Sigma>0$)	1	1	0 :)
d2 (1,2,-2)	0	0.2	1.1	-1.8	-1 ($\Sigma\leq 0$)	0	0	0 :)
d3 (1,-1,-1.5)	0	0.2	1.1	-1.85	-1 ($\Sigma\leq 0$)	0	0	0 :)
d4 (1,-2,-1)	0	0.2	1.1	-1.5	-1 ($\Sigma\leq 0$)	0	0	0 :)
d5 (1,-2,1)	0	0.2	1.1	0.7	1 ($\Sigma>0$)	1	1	0 :)
d6 (1,1.5,-0.5)	0	0.2	1.1	-0.25	-1 ($\Sigma\leq 0$)	0	1	1 :(

Updating weights for instance d6 (1, 1.5, -0.5) :

$w_0 = w_0 + \eta(y_6 - \hat{y}_6)x_0 = 0 + 0.2(1)(1) = 0.2$
$w_1 = w_1 + \eta(y_6 - \hat{y}_6)x_1 = 0.2 + 0.2(1)(1.5) = 0.2 + 0.3 = 0.5$
$w_2 = w_2 + \eta(y_6 - \hat{y}_6)x_2 = 1.1 + 0.2(1)(-0.5) = 1.1 - 0.1 = 1.0$

Instance d_k	w_0	w_1	w_2	Σ	Activation (Σ)	$\hat{y}_k$	y_k	Error/Cost
d1 (1,1,1)	0.2	0.5	1.0	1.7	1 ($\Sigma>0$)	1	1	0 :)
d2 (1,2,-2)	0.2	0.5	1.0	-0.8	-1 ($\Sigma\leq 0$)	0	0	0 :)
d3 (1,-1,-1.5)	0.2	0.5	1.0	-1.8	-1 ($\Sigma\leq 0$)	0	0	0 :)
d4 (1,-2,-1)	0.2	0.5	1.0	-1.8	-1 ($\Sigma\leq 0$)	0	0	0 :)
d5 (1,-2,1)	0.2	0.5	1.0	0.2	1 ($\Sigma>0$)	1	1	0 :)
d6 (1,1.5,-0.5)	0.2	0.5	1.0	0.45	1 ($\Sigma>0$)	1	1	0 :)

∴ Final weights learnt are : $w_0 = 0.2$, $w_1 = 0.5$ & $w_2 = 1.0$

Approximate target fn, $\hat{y} = 0.2 + 0.5x_1 + x_2$

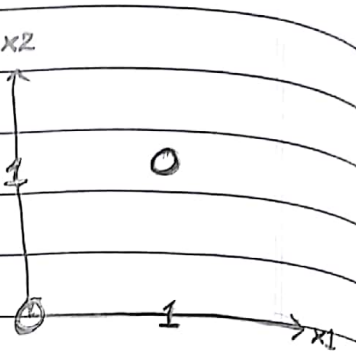
with initial $w_0 = 0$, $w_1 = 1$ & $w_2 = 0.5$ & $\eta = 0.2$

DATE: PAGE:

Why XOR gate can't be modeled using a single perceptron?

XOR Truth Table

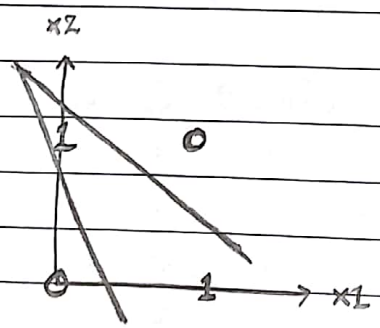
Instance	X1	X2	True label
d1	0	0	0
d2	0	1	1
d3	1	0	1
d4	1	1	0



Algorithmically, if we will be able to say that this dataset can not be separated by a linear fn when you feed this data as i/p to a single perceptron & the algorithm will never converge (infinite loop).

on plotting we find that the data (class 0 & 1) is not linearly separable. ∴ A single perceptron can't be used.

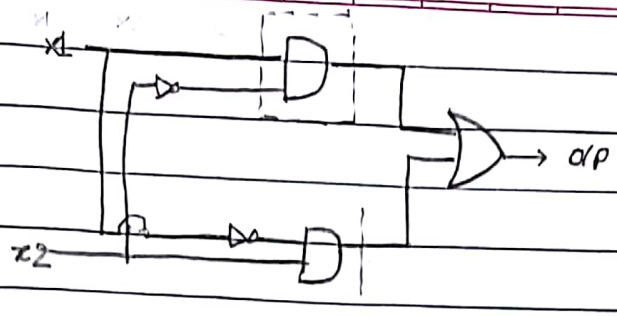
We would need more than one linear classifier (maybe a pair of lines)



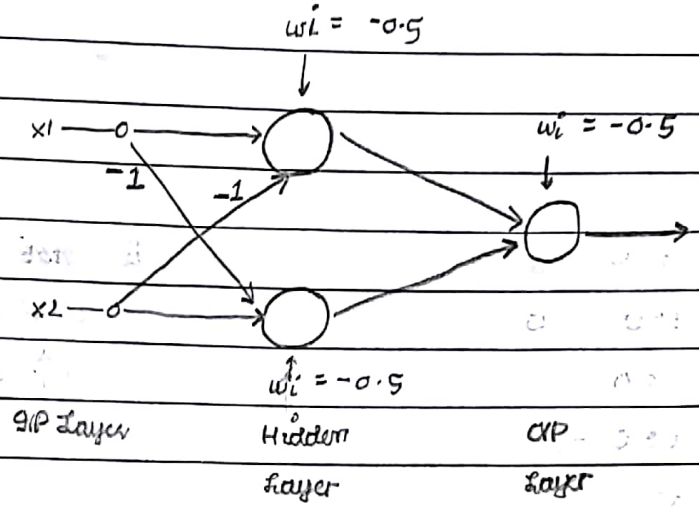
Modeling XOR Gate using more than one perceptron:

XOR Gate:

DATE: PAGE:



XOR Gate = $x_1x_2' + x_1'x_2$



14th August, 2017

Gradient Descent with Least Mean Square : DATE: PAGE:

For a particular board state b',

$$\hat{V}(b') = \hat{w}_{0b'} + \hat{w}_{1b'} x_{b'}$$

where

$x_{b'}$ is the # of black pawns on the board at board state b'.

For intermediate board state b',

$V_{train}(b')$ will not be given directly.

$\therefore V_{train}(b') = \hat{V}$ (Successor board state of b')

Error for this intermediate board state b',

$$E_{b'} = V_{train}(b') - \hat{V}(b')$$

Since b' is an intermediate board state b'

$$E_{b'} = \left[\hat{V}(\text{Successor board state of } b') - \hat{V}(b')\right]^2$$

$$\left[\text{value of } [\hat{w}_{0(b'+1)} + \hat{w}_{1(b'+1)} x_{1(b+1)}] - [\hat{w}_{0b'} + \hat{w}_{1b'} x_{1b'}]\right]^2$$

DATE: Intermediate PAGE:

We use Error calculated at board state b' to update $\hat{w}_{0b'}$ & $\hat{w}_{1b'}$ such that $w_{0b'}$ & $w_{1b'}$ attain the ideal value.

$$\hat{w}_{0b'} = \hat{w}_{0b'} - \eta E_{b'}$$

↓ This becomes equal to or close to equal to $w_{0b'}$

↓ Update this such that

$$= \hat{w}_{0b'} - \eta \left[\text{Value of } \left[\hat{w}_{0(b'+1)} + \hat{w}_{1(b'+1)} x_{1(b'+1)} \right] - \left(\hat{w}_{0b'} + \hat{w}_{1b'} x_{1b'} \right) \right]^2$$

$w_{0b'} \simeq \hat{w}_{0b'}$ when

$$\eta \left[\text{value of } \left[\hat{w}_{0(b'+1)} + \hat{w}_{1(b'+1)} x_{1(b'+1)} \right] - \left[\hat{w}_{0b'} + \hat{w}_{1b'} x_{1b'} \right] \right]^2 \simeq 0$$

$$\Rightarrow \eta E_{b'} \simeq 0$$

$E_{b'}$ is minimum wrt $\hat{w}_{0b'}$ when

$$\frac{dE_{b'}}{d\hat{w}_{0b'}} = 0$$

$$E_b' = (\text{value} - (\hat{w}_{0b}' + \hat{w}_{1b}' x_{1b}'))^2$$

DATE: PAGE:

$$\frac{dE_b'}{dw_{0b}'} = 2E_b'(0 - 1 + 0) = -2E_b'$$

For w_{1b}'

$$\frac{dE_b'}{dw_{1b}'} = 2E_b'(0 - (0 + x_{1b}'))^2 = -2E_b' x_{1b}'$$

So,

$$\hat{w}_{0b}' = \hat{w}_{0b}' - \eta E_b'$$

$$\hat{w}_{1b}' = \hat{w}_{1b}' - \eta E_b'$$

$\hat{w}_{0b}'$ becomes $\simeq w_{0b}'$ when E_b' is min. E_b' is min when wrt w_{0b}'

$$w_{0b}' = \hat{w}_{0b}' - \eta(-2E_b')$$

$$= \hat{w}_{0b}' + 2\eta E_b'$$

$\because 2$ is a constant

$$2\eta \simeq \eta$$

$$\boxed{w_{0b}' = \hat{w}_{0b}' + \eta E_b'}$$

&

Similarly for w_{1b}' we get,

$$w_{1b}' = \hat{w}_{1b}' - \eta(-2E_b' x_{1b}')$$

$$= \hat{w}_{1b}' + 2\eta E_b' x_{1b}'$$

Considering $2\eta \simeq \eta$

$$w_{16'} = \hat{w}_{16'} + \eta E_{6'} x_{16'}$$

DATE: PAGE:

$$w_{16'} = \hat{w}_{16'} + \eta \left[V_{train}(6') - \hat{V}(6) \right] x_{16'}$$

DATE: PAGE:

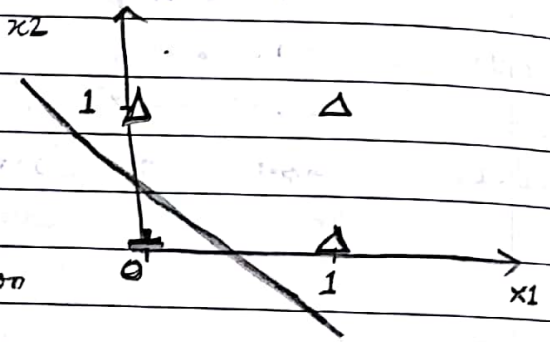
A single perceptron and OR Gate :

Truth Table For OR Gate

Instance	x1	x2	True Label
d1	0	0	0 (+)
d2	0	1	1 (Δ)
d3	1	0	1 (Δ)
d4	1	1	1 (Δ)

Is the data linearly separable ?

Yes. The data is linearly separable.



∴ If a single perceptron is used, we'll be able to classify the data (learning algorithm : gradient descent will converge)

Let initial values for $w_1 = 1$, $\eta = 0.2$

$w_2 = -0.5$, $w_0 = -0.5$, $x_0 = 1$

Target fn representation, $\hat{y} = w_0 + w_1 x_1 + w_2 x_2$

Instance $d_k(x_0, x_1, x_2)$	w_0	w_1	w_2	Σ	Activation	$\hat{y}$	y	Cost/Error (Measure of happiness)
d1 (1, 0, 0)	-0.5	1	0.5	-0.5	-1 ($\Sigma \leq 0$)	0	0	0 :)
d2 (1, 0, 1)	-0.5	1	0.5	0	-1 ($\Sigma \leq 0$)	0	1	1 :(

Updating weights with instance d2 (1, 0, 1)

$w_0 = w_0 + \eta$ (True Label − Predicted Label) x_0

$= -0.5 + 0.2\,(1)(1)$

$= -0.5 + 0.2 = -0.3$

Instance d_k (x_0, x
d1 (1, 0, 0)
d2 (1, 0, 1)
d3 (1, 1, 0)
d4 (1, 1, 1)

The

Appro

DATE: PAGE:

$w_1 = w_1 + \eta$ (True Label − Predicted Label) x_1

$= 1 + 0.2(1)(0)$

$= 1$

$w_2 = w_1 + \eta$ (True Label − Predicted Label) x_1

$= 0.5 + 0.2(1)(1)$

$= 0.5 + 0.2$

$= 0.7$

Instance	$d_k(x_0, x_1, x_2)$	w_0	w_1	w_2	Σ	Activation(Σ)	$\hat{y}$	y	Cost/Error
d_1	(1, 0, 0)	−0.3	1	0.7	−0.3	$-1 (\Sigma \leq 0)$	0	0	0 :)
d_2	(1, 0, 1)	−0.3	1	0.7	0.4	$+1 (\Sigma > 0)$	1	1	0 :)
d_3	(1, 1, 0)	−0.3	1	0.7	0.7	$+1 (\Sigma > 0)$	1	1	0 :)
d_4	(1, 1, 1)	−0.3	1	0.7	1.4	$+1 (\Sigma > 0)$	1	1	0 :)

The algorithm converged with

$w_0 = -0.3$

$w_1 = 1$

$w_2 = 0.7$

Approximated target fn, $\hat{y} = -0.3 + x_1 + 0.7 x_2$

Started with initial $w_0 = -0.5$

$w_1 = 1$

$w_2 = 0.5$

& $\eta = 0.2$

Label
(+)
(Δ)
(Δ)
(Δ)

(Degree of happiness)

:)
:(

Candidate Elimination DATE: PAGE:

Algorithm

Starting with

-ve Training Data

Training Data

Instance	Size	Color	Shape	Class / Label
d1	big	red	circle	No
d2	small	red	triangle	No
d3	small	red	circle	Yes
d4	big	blue	circle	No
d5	small	blue	circle	Yes

Soi: So $< \phi \ \phi \ \phi >$

G_0 $< ? \ ? \ ? >$

When you are presented with a negative example, you need to remove from S any hypothesis inconsistent with the current observation and replace any hypothesis in G with its minimal specializations (all) that are consistent with the observation but still more general than some member of S.

Considering d1 < big red circle No >

-ve example,

1. Consistent with all hypothesis in So ?

   Yes ⇒ Do nothing

2. consistent with all " " G_0 ?

   No ⇒ i) Replace the inconsistent hypothesis in G_0 with all its minimal specialisations that are

$G_1 = \{ <$ small ? ?>,

$<$? , blue ?>,

$<$? ? triangle $>\}$

DATE: PAGE:

ii) Remove from G1 any hypothesis that is less general than another " in G1
No such hypothesis.

∴ After d1,

S1 = S0 = < ϕ ϕ ϕ >

G1 = { < small ? ? >,
< ? blue ? >,
< ? ? triangle > }

Instance d2,

d2 < small red triangle No >

Since its a -ve example,

1. Consistent with S1 ? Yes ⇒ Do nothing S2 = S1
2. " " all hypothesis in G1 ?
No. It is not consistent with ($h(d_k) \neq c(d_k)$)
< small ? ? > { h(d2) = 1, c(d2) = 0 }
and
< ? ? triangle > { h(d2) = 1, c(d2) = 0 }

i) Replace < small ? ? >
with all its minimal specialisations
such that they are consistent with the observation
(instance d2) but still more general than
some member of S2.
Do the same for < ? ? triangle >.

are consistent with the observation but still more general than
some member of S

DATE: PAGE:

Doing it for

< small ? ? >

we get

< small blue ? > ✓ , < small ? circle > ✓

For < ? ? triangle >

we get

< big ? triangle > ✓ < ? blue triangle >

∴ G2 = { < ? blue ? > ,

< small blue ? > , < small ? circle >

< big ? triangle > < ? blue triangle > }

ii) Remove from G2 any hypothesis which is less general than any other hypothesis in G2

< small blue ? > & < ? blue triangle > is less general than

< ? blue ? >

∴ G2 = { < ? blue ? > < small ? circle >

< big ? triangle > }

After instance d2,

S2 = < ϕ ϕ ϕ >

G2 = { < ? blue ? > < small ? circle >

< big ? triangle > }

Instance d3, < small red circle Yes >

DATE: PAGE:

For +ve example,

1. Consistent with all hypothesis in G2 ? No.
   Remove the inconsistent hypotheses from G2.

   G3 = { < small ? circle > }

2. Consistent with all hypothesis in S2 ? No

   i) Replace with all its (inconsistent hypothesis) minimal generalisations but still it should be more specific than some member of G3.

   S3 = < small red circle >

   ii) Remove from S3 any hypothesis that is more general than any other hypothesis in S3.

   S3 = < small red circle >

After instance d3,

G3 = { < small red circle > }
G3 = { < small ? circle > }

Similarly doing it for d4 < big blue circle No >

S4 = { < small red circle > }
G4 = { < small ? circle > }

Doing it for d5 < small blue circle Yes > we get:

S5 = { < small ? circle > }
G5 = { < small ? circle > }

∵ S5 = G5 = < small ? circle >

DATE: PAGE:

Approximation of target concept = < small ? wide ?

Finally we have learnt the concept of "small circles"

Derivation of the Back propagation Rule :

$$E = \frac{1}{2} \sum_{k=1}^{d} (y_k - \hat{y}_k)^2$$

where

$$\hat{y}_k = w_0 + \sum_{i=1}^{n} w_i x_i$$

Considering $\hat{y}_k = w_0 + w_1 x_1$

then,

$$E = \frac{1}{2} \sum_{k} (y_k - (w_0 - w_1 x_1))^2$$

Update rule,

w

E on the entire training dataset,

$$E = \frac{1}{2} \sum_{k=1}^{\#\text{ of examples}} (y_k - \hat{y}_k)^2$$

For a particular training instance d_k

$$E_k = \frac{1}{2} (y_k - \hat{y}_k)^2$$

$$= \frac{1}{2} (y_k - (w_0 + w_1 x_1))^2$$

DATE: PAGE:

c>

To update the weight, for instance k

$$w_i' = w_i - \eta \frac{\partial E_k}{\partial w_i}$$

$$w_0 = w_0 - \eta \frac{\partial}{\partial w_0}\left(\frac{1}{2}(y_k - (w_0 + w_1 x_1))^2\right)$$

$$= w_0 - \frac{1}{2}\eta\left(2 \times (y_k - \hat{y_k}) \times (0 - 1)\right)$$

$$= w_0 + \eta (y_k - \hat{y_k})$$

$$w_1 = w_1 - \eta \frac{\partial}{\partial w_1}\left(\frac{1}{2}(y_k - (w_0 + w_1 x_1))^2\right)$$

$$= w_1 - \eta\, 2 \times \frac{1}{2} \times (y_k - \hat{y_k})(0 - (0 + x_1))$$

$$= w_1 + \eta (y_k - \hat{y_k})(x_1)$$

T2 DATE: PAGE:

Gradient Descent :

$$E = \frac{1}{2} \sum_{d \in D} (\text{True Label} - \text{Predicted Label})^2$$

Perceptron Update Rule : Discrete Value ∵ Step Fn is used

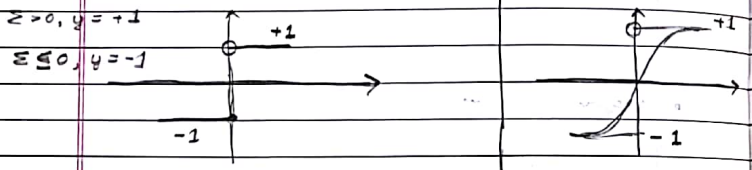
- $w_i \leftarrow w_i + \eta (\text{True label} - \text{Predicted Label}) x_i^d$

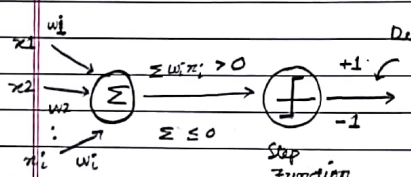
Gradient Descent Update Rule :

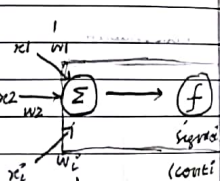
$$w_i \leftarrow w_i + \eta \sum_{d \in D} (\text{True Label} - \text{Predicted Label}) x_i^d$$

General Gradient Descent

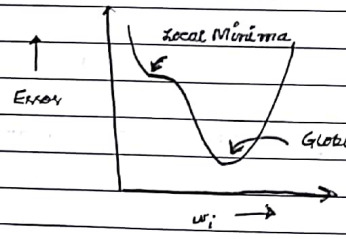
- Step Function : Non Differentiable
- Sigmoid Fn : Differentiable, its a continuous fn







Momentum :



To prevent from getting stuck at local minima.

$\Delta w_{ji} \leftarrow \eta \delta_j x_{ji} + \alpha w_{ji}(n-1)$

Momentum Fn,

$0 \le \alpha < 1$

1) Step fn vs Sigmoid fn

DATE: PAGE:

Single vs Neural Perceptron Network

Linearly vs Non Separable Linearly Separable

2) Momentum

- Differentiation of $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\frac{d}{dx}(\sigma(x)) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

$$= -\frac{1}{(1+e^{-x})^2}(0 - e^{-x})$$

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

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

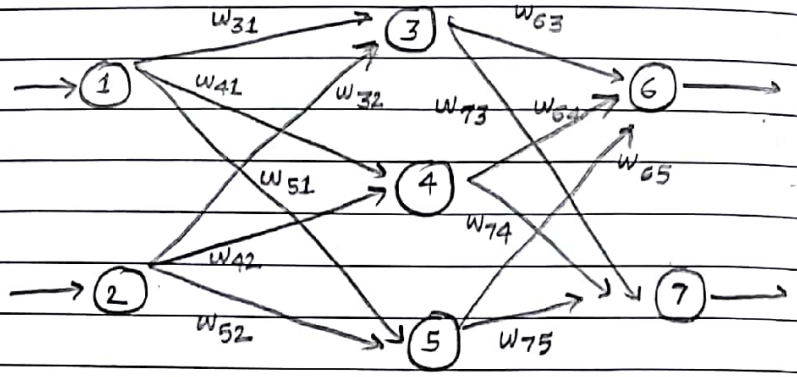
$$= \sigma(x) - \sigma(x)^2$$

$$= \sigma(1-\sigma)$$

DATE: PAGE:

$$\frac{d\sigma}{dx} = \sigma(1-\sigma)$$

Neural Network : # of o/p from hidden layer < # of o/p from o/p layer
∴ This is "Downsizing"



k - o/p neurons

I/P Layer	Hidden layer (Each neuron in the hidden layer is a Sigmoid unit)	Output layer (Each neuron in the hidden layer is a Sigmoid unit)

where each sig

We can update the weight of the neural network either by using General Gradient Descent Rule

or

by using Stochastic Gradient Descent Rule.

Though, Stochastic Gradient Descent Rule gives better accuracy generally.

General Gradient Descent Rule

$$w_i \leftarrow w_i + \eta \frac{1}{2} \sum_{d \in D} (\text{True Label} - \text{Predicted Label}) x_i$$

Stochastic Gra

$w_i \leftarrow$

we do

Back propagation

$E_d(\vec{w})$

Δw_{ji}

DATE: PAGE:

Stochastic Gradient Descent

$w_i \leftarrow w_i + \eta$ (True Label - Predicted Label) x_i

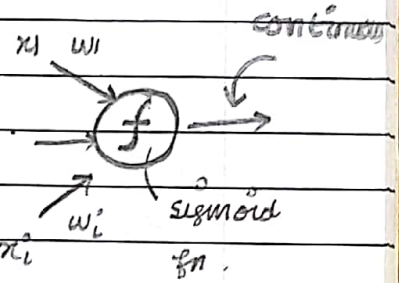
we do this after every example.

o/p from o/p layer

∴ This is "Downstream"

neurons -

where each sigmoid unit is



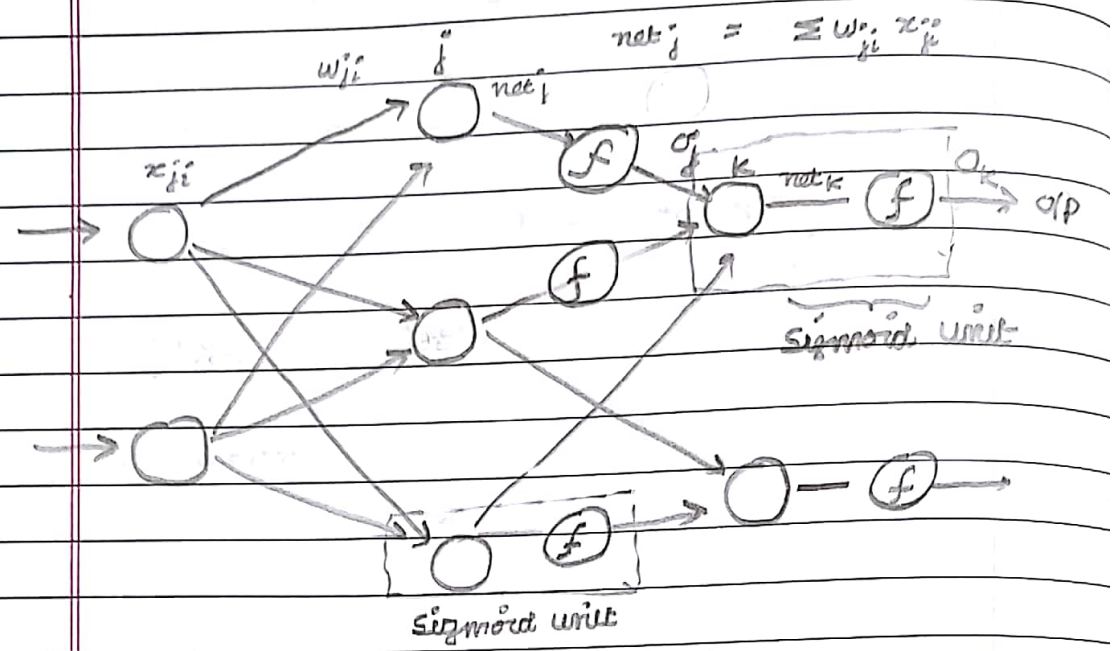
$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Back-propagation Algorithm for Output Layer to Update weights:

$$E_d(\vec{w}) = \frac{1}{2} \sum_{d \in D} \sum_{k \in \text{\# of labels/classes}} (t_{kd} - o_{kd})^2$$

$$\boxed{\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}}} \quad - (1B)$$

Nomenclature of Neural Network : DATE: PAGE:



$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \times \frac{\partial net_j}{\partial w_{ji}}$$

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \times x_{ji} \quad \text{— (1A)}$$

Case 1 : Training rule for O/P units

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j} \times \frac{\partial o_j}{\partial net_j}$$

$$= \frac{1}{2} \times 2 \times \sum_{k \in \text{\# of Cabels}} (t_k - o_k)^2$$

$$= \frac{1}{2} \times 2 (t_j - o_j) \times \frac{\partial}{\partial o_j} (t_j - o_j)$$

$$= (t_j - o_j)(-1)$$

DATE: PAGE:

$$\frac{\partial E_d}{\partial o_j} = -(t_j - o_j) \quad -(2)$$

Since $o_j = \sigma(net_j)$

Now, 2nd term in $\frac{\partial o_j}{\partial net_j} = \frac{\partial(\sigma(net_j))}{\partial net_j}$

$$\frac{\partial o_j}{\partial net_j} = \sigma(net_j)[1 - \sigma(net_j)]$$

Substituting $\sigma(net_j) = o_j$

$$\frac{\partial o_j}{\partial net_j} = o_j(1 - o_j) \quad -(3)$$

Substitute (4) in (1A) we get

$$\frac{\partial E_d}{\partial w_{ji}} = -(t_j - o_j) o_j (1 - o_j) x_{ji} \quad -(5)$$

Substituting eqn (5) in eqn (1B)

$$\Delta w_{ji} = \eta (t_j - o_j) o_j (1 - o_j) x_{ji}$$

3rd October, 2017

DATE: PAGE:

$$\Delta w_{ji} = \Delta w_{ji} + \eta (t_j - o_j) o_j (1 - o_j) x_{ji}$$

$$\frac{\partial E}{\partial net_j} = \sum_{k \in ds(j)} \frac{\partial E}{\partial net_k} \cdot \frac{\partial net_k}{\partial net_j}$$

We represent, $\frac{\partial E}{\partial net_k} = -\delta_k$

$$\frac{\partial E}{\partial net_j} = \sum_{k \in ds(j)} -\delta_k \cdot \frac{\partial net_k}{\partial net_j}$$

$$= \sum_{k \in ds(j)} -\delta_k \cdot \frac{\partial net_k}{\partial o_j} \cdot \frac{\partial o_j}{\partial net_j}$$

Rewriting, $\frac{\partial net_k}{\partial o_j} = w_{ji}$

we get :

$$\sum -\delta_k w_{ji} \frac{\partial o_j}{\partial net_j}$$

$$net_j = \sum w_j x_j = w_0 x_0 + w_1 x_1 + \cdots +$$

$$\frac{\partial net_j}{\partial x_j} = w_0 + w_1 + \cdots + \cdots = \sum w_j$$

k-cross Validation : used when # of training DATE: examples PAGE: is less

i) 100 examples.

ii) Divide 100 examples into k-set of 25

iii) Some mean.

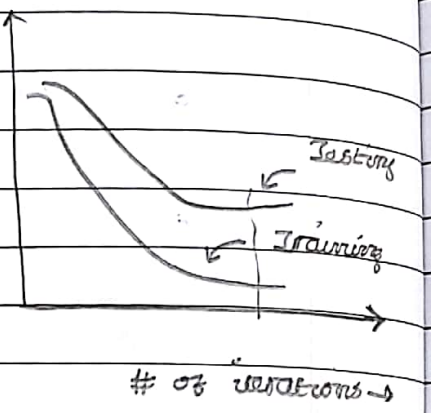
Calculating ideal # of iterations:

60% Training

40% Validation

Stored wt. w_d | w_n

Best fit w_0 | w_n



Bias in Neural Network:

We assume that there exists a smooth surface to fit the data set.

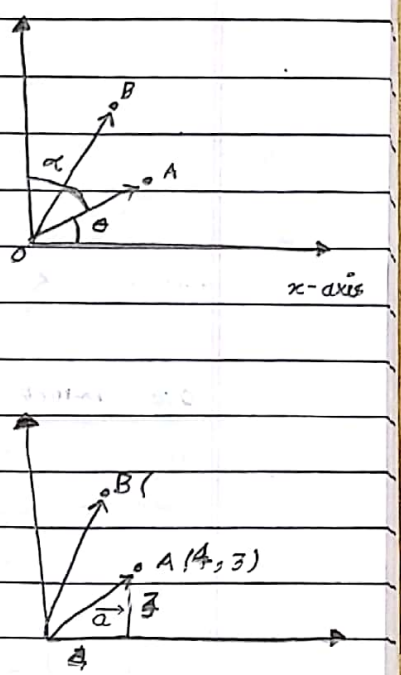
Magnitude of $\vec{OA}$ = Norm of $\vec{OA}$

$= \sqrt{(4-0)^2 + (3-0)^2}$

$= \sqrt{16+9}$

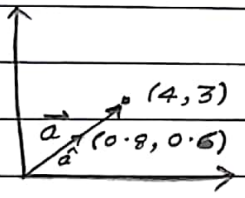
$= \sqrt{25}$

$= 5$



Direction of $\vec{OA}$ (or $\vec{a}$) = $(\cos\theta, \cos\alpha)$

= [cos (angle made with horizontal axis), cos (angle made with vertical axis)]



$= \left[\frac{4}{5}, \frac{3}{5}\right] = [0.8, 0.6]$

$= \left[\frac{a1}{\sqrt{a1^2+a2^2}}, \frac{a2}{\sqrt{a1^2+a2^2}}\right]$

$\hat{a} = \left[\frac{a1}{\|\vec{a}\|}, \frac{a2}{\|\vec{a}\|}\right]$

NOTE : Direction vector of $\vec{a}$ is a unit vector.

$\hat{a} = [0.8, 0.6]$

$\|\hat{a}\| = \sqrt{(0.8-0)^2 + (0.6-0)^2}$

$= \sqrt{0.8^2 + 0.6^2}$

$= \sqrt{0.64 + 0.36}$

$= \sqrt{1.0}$

$= 1$

DATE: PAGE:

Given a vector $\vec{a}\ (a1, a2)$, calculate :

i) Magnitude / norm of $\vec{a}$,

$$\|\vec{a}\| = \sqrt{a1^2 + a2^2}$$

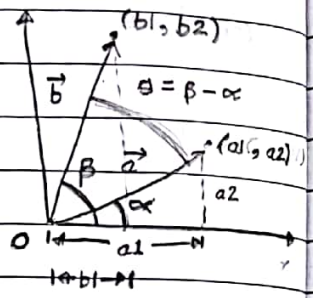
ii) Unit vector of $\vec{a}$,

$$\hat{a} = \left(\frac{a1}{\|\vec{a}\|}, \frac{a2}{\|\vec{a}\|} \right)$$

Also, $\vec{a} = \|\vec{a}\| \hat{a}$

Dot Product :

$$\cos\alpha = \frac{a1}{\|\vec{a}\|}, \quad \cos\beta = \frac{b1}{\|\vec{b}\|}$$



$$\cos(\beta - \alpha) = \cos\theta$$

$$= \cos\beta \cos\alpha + \sin\beta \sin\alpha$$

$$= \frac{b1}{\|\vec{b}\|} \frac{a1}{\|\vec{a}\|} + \frac{b2}{\|\vec{b}\|} \frac{a2}{\|\vec{a}\|}$$

$$\cos\theta = \frac{b1\,a1 + b2\,a2}{\|\vec{b}\|\|\vec{a}\|}$$

$$\cos\theta \|\vec{b}\| \|\vec{a}\| = b1\,a1 + b2\,a2 = \text{Product of 1st co-ordinate of } \vec{a} \text{ \& } \vec{b} + \text{Product of 2nd co-ordinate of } \vec{a} \text{ \& } \vec{b}$$

$$= \sum b_i a_i$$

= Dot Product

or

Inner Product

Projection o

Projection

is

in the

Direction

Projection

Examples :

$\hat{w} =$ (iw

Projection

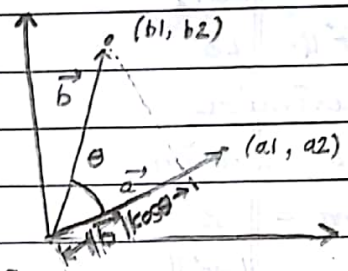
DATE: PAGE:

Projection of a vector on another vector :

Projection of $\vec{b}$ on $\vec{a}$

is

in the direction of $\vec{a}$



Direction of $\vec{a}$, $\hat{a} = \left[\frac{a1}{\|\vec{a}\|}, \frac{a2}{\|\vec{a}\|}\right]$

Projection of $\vec{b}$ on $\vec{a}$ = $\underbrace{\|\vec{b}\|\cos\theta}_{\text{Magnitude}} \cdot \underbrace{\hat{a}}_{\text{Direction}}$

$\vec{b}$

$\|\vec{a}\| \times \|\vec{b}\| \cos\theta \; \frac{\vec{a}}{\|\vec{a}\|}$

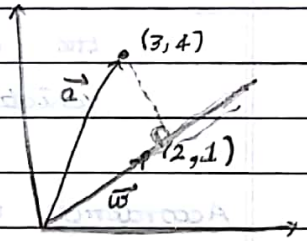
$\frac{}{\|\vec{a}\|} \qquad \frac{}{\|\vec{a}\|}$

$\frac{(\vec{a}\cdot\vec{b})\,\vec{a}}{\|\vec{a}\|}$

Example : Find projection of

$\vec{a}$ on $\vec{w}$

$\hat{w} = \left(\frac{w1}{\|\vec{w}\|}, \frac{w2}{\|\vec{w}\|}\right) = \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right)$



oduct of 2nd

o-ordinate of $\vec{a}$ & $\vec{b}$

$\|$ Projection of $\vec{a}$ on $\hat{w}\|$ $= \|\vec{a}\|\cos\theta$

$= \frac{\|\vec{a}\|\|\hat{w}\|\cos\theta}{\|\hat{w}\|}$

$= \frac{\vec{a}\cdot\hat{w}}{\|\hat{w}\|}$

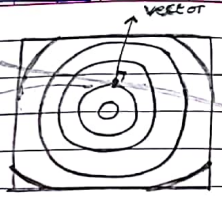
$= \frac{3\left(\frac{2}{\sqrt{5}}\right) + 4\left(\frac{1}{\sqrt{5}}\right)}{1}$

$= \frac{10}{\sqrt{5}}$

Support Vector Machine

DATE: PAGE:

Minimizing $\|\vec{w}\|$ to maximize the margin $m = \frac{2}{\|\vec{w}\|}$



Critical Point: where constraint line is tangent to the circle

Concentric circles: $x^2 + y^2 = r^2$

using Lagrange Multiplier:

Objective is

$$f(w,b) = \frac{\|\vec{w}\|^2}{2}$$ — Primal Function

constraint

$$g(w,b) = y_i(\vec{w}\vec{x} + \vec{b}) - 1 \geq 0$$

Solution: Note that we are trying to minimize the value of $\|\vec{w}\|$. $\therefore$ $\vec{w}$ is a variable.

According to Lagrange,

$$f(w,b) = \alpha\,[g(w,b)]$$

where

α: Lagrange Multiplier

Lagrange representation in the form of above eqn:

$$\Rightarrow \quad f(w,b) - \alpha g(w,b) = 0$$

$$L(w,b,\alpha) = \nabla_{(w,b)} f(w,b) - \alpha \nabla_{(w,b)} g(w,b) = 0$$

DATE: PAGE:

the line $y = \vec{w}x + b$

we know that $\vec{w}$ is also $\perp$ to the line $\vec{w}x + b$

$\therefore$ $\vec{w}$ & the vector $\perp$ to the point will be parallel.

$$L(w,b,\alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i(w_i x_i + b) - 1\right) = 0$$

$$= \frac{w_i^2}{2} - \sum_{i=1}^{m} \alpha_i \left(y_i(w_i x_i + b) - 1\right) = 0 \quad (1)$$

$$= \frac{w_i^2}{2} - \sum_{i=1}^{m} \alpha_i y_i (w_i x_i + b) + \sum_{i=1}^{m} \alpha_i = 0$$

Finding Change in expression (1) wrt w

$$\nabla_w L(w,b,\alpha) = \nabla_w f - \alpha \nabla_w g$$

where $f = \frac{1}{2}\|\vec{w}\|^2 = \frac{w_i^2}{2}$

$$g = \alpha_i \sum_{i=1}^{m} y_i (w_i x_i + b) - 1$$

$$\nabla_w f = \frac{\partial f}{\partial w} = w_i$$

$$\nabla_w g = \frac{\partial g}{\partial w} = \alpha_i \sum_{i=1}^{m} y_i x_i$$

$$\therefore \quad \boxed{\nabla_w L(w,b,\alpha) = w_i - \alpha_i \sum_{i=1}^{m} y_i x_i} \quad - (2)$$

$$w_i - \alpha_i \sum_{i=1}^{m} y_i x_i = 0$$

$$\Rightarrow w_i = \alpha_i \sum_{i=1}^{m} y_i x_i$$

Finding Change in expression (1) wrt b

$$\nabla_b L(w,b,\alpha) = \nabla_b f - \alpha \nabla_b g$$

where $f = \frac{\|\vec{w}\|^2}{2} = \frac{w_i^2}{2}$

$$g = \alpha_i \sum_{i=1}^{m} \left[y_i(w_i x_i + b) - 1\right]$$

DATE: PAGE:

$\nabla_b f = \frac{\partial f}{\partial b} = 0$, $\nabla_b g = \frac{\partial g}{\partial b}$

$= \frac{\partial}{\partial b}\left(\alpha_i \sum_{i=1}^{m} [y_i(w_i x_i + b) - 1]\right)$

$= 0 \; \alpha_i \sum_{i=1}^{m} [y_i]$

$= \alpha_i \sum_{i=1}^{m} y_i$

$\therefore \nabla_b L = \nabla_b f - \nabla_b g = 0$

$= 0 - \alpha_i \sum_{i=1}^{m} y_i = 0$

$\Rightarrow \boxed{\alpha_i \sum_{i=1}^{m} y_i = 0} - (3)$

From (1),

$L(w, \alpha, b) = \frac{w_i^2}{2} - \alpha_i \sum_{i=1}^{m} [y_i(w_i x_i + b) - 1] = 0$

$\Rightarrow = \frac{w_i^2}{2} - \alpha_i \sum_{i=1}^{m} y_i w_i x_i - \alpha_i \sum_{i=1}^{m} y_i b - \alpha_i \sum_{i=1}^{m} 1$

Using (2), $w_i = \sum_{i=1}^{m} \alpha_i y_i x_i$

" (3), $\alpha_i \sum_{i=1}^{m} y_i = 0$

$= \frac{\left[\sum_{i=1}^{m} \alpha_i y_i x_i\right]^2}{2} - \alpha_i \sum_{i=1}^{m} y_i w_i x_i - b\alpha_i$

$= \frac{1}{2}\left[\sum_{i=1}^{m} \alpha_i y_i x_i\right]^2 - \alpha_i \sum_{i=1}^{m} y_i w_i x_i - 0 -$

$\Rightarrow \theta($

θ using (3)

$\sum_{i=1}^{m} y_i - \sum_{i=1}^{m} \alpha_i$

$\sum_{i=1}^{m} \alpha_i$

DATE: PAGE:

$$\Rightarrow \theta(\alpha) = \frac{1}{2}\left[\sum_{i=1}^{m} \alpha_i y_i x_i\right]^2$$

$$-\left[\sum_{i=1}^{m} \alpha_i y_i x_i \sum_{j=1}^{m} \alpha_j y_j x_j - \sum_{i=1}^{m} \alpha_i\right]$$

$$\left\{\because w_i = \sum_{i=1}^{m} \alpha_i y_i x_i\right\}$$

$$= \frac{1}{2}\left[\sum_{i=1}^{m} \alpha_i y_i x_i \sum_{j=1}^{m} \alpha_j y_j x_j\right]$$

$$-\sum_{i=1}^{m} \alpha_i y_i x_i \sum_{j=1}^{m} \alpha_j y_j x_j$$

$$+\sum_{i=1}^{m} \alpha_i$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\left[\sum_{i=1}^{m} \alpha_i y_i x_i \sum_{j=1}^{m} \alpha_j y_j w_j\right]$$

$$- \alpha_i \sum_{i=1}^{m} 1$$

$$\boxed{\theta(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j}$$

$\Downarrow$

θ using (3)

$$\sum_{i=1}^{m} y_i - \sum_{i=1}^{m} \alpha_i$$

$$\sum_{i=1}^{m} \alpha_i$$

Done using Sequential Minimization Optimization Algo (SMO Algo)

- We need to maximize $\theta(\alpha)$
- to minimize $\|\vec{w}\|$ which maximizes margin m.

→ Out of Syllabus.

DATE: PAGE:

Problem :

Consider the sample data :

	a1	a2	True Label (y_i)
x1	1	3	+1 (y_1)
x2	3	1	-1 (y_2)

Find optimal plane using SVM to separate the data.

Sol: Using eqn 3,

$$\sum_{i=1}^{2} \alpha_i y_i = 0$$

$\therefore$ Here, $\sum_{i=1}^{2} \alpha_i y_i = 0$

$$\alpha_1 y_1 + \alpha_2 y_2 = 0$$

$$\alpha_1 (+1) + \alpha_2 (-1) = 0$$

$$\alpha_1 - \alpha_2 = 0$$

$$\boxed{\alpha_1 = \alpha_2} \quad -(1)$$

$$\theta(\alpha) = \sum_{i=1}^{2} \alpha_i - \frac{1}{2} \left[\sum_{i=1}^{2} \sum_{j=1}^{2} \alpha_i \alpha_j y_i y_j x_i x_j \right]$$

$$= [\alpha_1 + \alpha_2] - \frac{1}{2} \left[\sum_{i=1}^{2} \right.$$

$\alpha_1 \alpha_1 y_1 y_1 x_1 x_1$

+

$\alpha_1 \alpha_2 y_1 y_2 x_1 x_2$

+

$\alpha_2 \alpha_1 y_2 y_1 x_2 x_1$

+

$\alpha_2 \alpha_2 y_2 y_2 x_2 x_2 \left. \right]$

$= (\alpha_1 +$

$= 2\alpha_1$

$= 2\alpha$

$= 2\alpha_1$

$= 2\alpha_1$

$\Rightarrow -4\alpha$

$\Rightarrow -4\alpha$

$\Rightarrow \alpha_1$

2

DATE: PAGE:

Using $\alpha_1 = \alpha_2$,

$$= (\alpha_1 + \alpha_2) - \frac{1}{2}\left[\alpha_1^2 y_1^2 x_1^2 + 2\alpha_1^2 y_1 y_2 x_1 x_2 + \alpha_1^2 y_2^2 x_2^2\right]$$

$$= 2\alpha_1 - \frac{1}{2}\Big[\alpha_1^2 (+1)^2 \underset{(1,3)(1,3)}{} + 2\alpha_1^2 (+1)(-1)\underset{(3,1)}{(1,3)(3,1)} + \alpha_1^2 (-1)^2 \underset{(3,1)}{(3,1)}\Big]$$

$$= 2\alpha_1 - \frac{1}{2}\left[\alpha_1^2 (1+9) + 2\alpha_1^2 (-1)(3+3) + \alpha_1^2 (9+1)\right]$$

$$= 2\alpha_1 - \frac{1}{2}\left[10\alpha_1^2 - 12\alpha_1^2 + 10\alpha_1^2\right]$$

$$= 2\alpha_1 - \frac{8\alpha_1^2}{2} = 0$$

$$\Rightarrow -4\alpha_1^2 + 2\alpha_1 = 0$$

$$\Rightarrow -4\alpha_1^2 = -2\alpha_1 \Rightarrow \alpha_1 = 2$$

$$2\alpha_1 - 4\alpha_1^2$$

$$2\alpha_1 [1 - 2\alpha_1] = 0$$

$$\Rightarrow 2\alpha_1 - 4\alpha_1^2 = 0$$

$$\Rightarrow 2\alpha_1 [1 - 2\alpha_1] = 0$$

$$\boxed{\alpha_1 = 0 \text{ or } \alpha_1 = 1/2}$$

2 roots are

x_i°]

$\alpha_1\alpha_1 y_1 y_1 x_1 x_1$
+
$\alpha_1\alpha_2 y_1 y_2 x_1 x_2$
+
$\alpha_2\alpha_1 y_2 y_1 x_2 x_1$
+
$\alpha_2\alpha_2 y_2 y_2 x_2 x_2$]

maximizing $(O(\alpha))$

$\Rightarrow$ Diff. wrt α

$$-8\alpha_1 + 2 = 0$$

$$\Rightarrow -8\alpha_1 = -2$$

$$\Rightarrow \boxed{\alpha_1 = 1/4}$$

DATE: PAGE:

At $\alpha_1 = 1/4$,

$$\theta(\alpha) = 2\left(\frac{1}{4}\right) - \frac{1}{2} \times 8 \times \left(\frac{1}{4}\right)^2$$

$$= \frac{1}{2} - \frac{4 \times 1}{4} \times \frac{1}{4}$$

$$= \frac{1}{2} - \frac{1}{4}$$

$$= -\frac{1}{4} < 0$$

$\therefore$ at $\alpha = 1/4$, $\theta(\alpha)$ is max.

$\because \alpha_1 = \alpha_2$

$\therefore \alpha_1 = \alpha_2 = 1/4$

Calculating w,

$$w = \sum y_i x_i \alpha_i$$

$$= y_1 x_1 \alpha_1 + \alpha_2 y_2 x_2$$

$$= (+1)(1,3)\frac{(1)}{4} + \left(\frac{1}{4}\right)(-1)(3,1)$$

$$= \left(\frac{1}{4}, \frac{3}{4}\right) - \left(\frac{3}{4}, \frac{1}{4}\right)$$

$$= \left(-\frac{2}{4}, \frac{2}{4}\right)$$

$$\vec{w} = \left(-\frac{1}{2}, \frac{1}{2}\right)$$

$$\|\vec{w}\| = \sqrt{\frac{1}{4} + \frac{1}{4}} = \sqrt{\frac{2}{4}} = \frac{1}{\sqrt{2}}$$

$$m = \frac{2}{\|\vec{w}\|} = \frac{2}{\left(\frac{1}{\sqrt{2}}\right)} = 2\sqrt{2}$$

DATE: PAGE:

Eqn of 2 hyperplanes are :

$$wx_1 + b = +1$$

$$+ \quad wx_2 + b = -1$$

$$wx_1 + wx_2 + 2b = 0$$

$$2b = -\left[\left(-\frac{1}{2}, \frac{1}{2}\right)(1,3) + \left(-\frac{1}{2}, \frac{1}{2}\right)(3,1)\right]$$

$$2b = -\left[\left(\frac{-1}{2} + \frac{3}{2}\right) + \left(\frac{-3}{2} + \frac{1}{2}\right)\right]$$

$$2b = -[1 - 1] = 0$$

$$b = 0$$

Hyperplane eqn : $wx + b = 0$

$$\Rightarrow (w1, w2)\begin{pmatrix} a1 \\ a2 \end{pmatrix} + 0 = 0$$

$$\Rightarrow \left(-\frac{1}{2}, \frac{1}{2}\right)(a1, a2) = 0$$

$$\Rightarrow \frac{-a1}{2} + \frac{a2}{2} = 0$$

$$\Rightarrow \boxed{a2 = a1}$$

Hyperplane 1 : $wx + b = +1$

$$\left(-\frac{1}{2}, \frac{1}{2}\right)(a1, a2) + 0 = +1$$

$$\frac{-a1}{2} + \frac{a2}{2} = +1$$

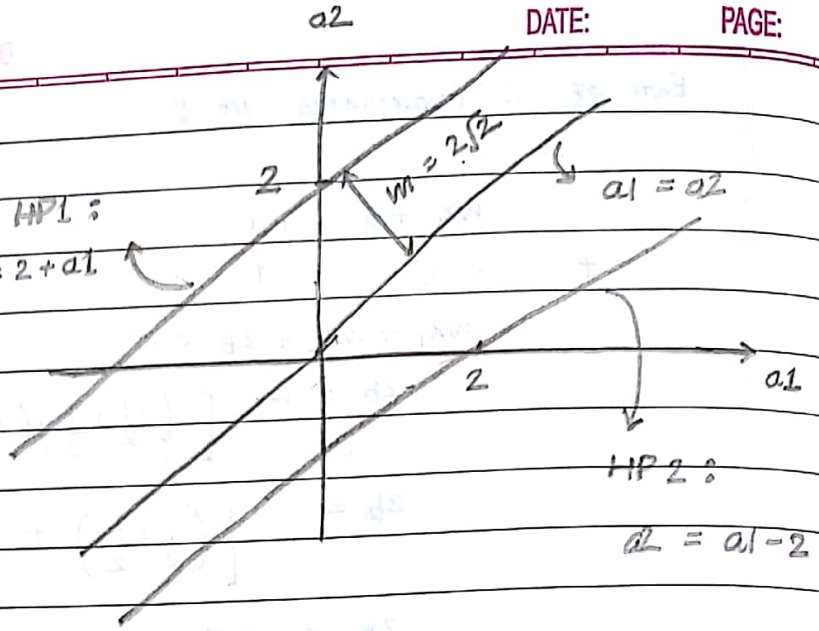
$$\boxed{a2 = 2 + a1}$$

" 2 : $wx + b = -1$

$$\left(-\frac{1}{2}, \frac{1}{2}\right)(a1, a2) + 0 = -1$$

$$\boxed{a2 = a1 - 2}$$

DATE: PAGE:



Genetic Algorithm

DATE: PAGE:

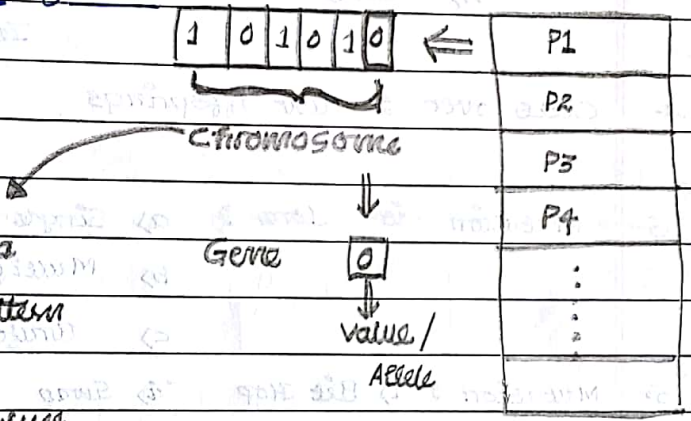
i) Search Based Technique.

ii) If Optimization is reqd., we use Genetic ,,

Application :
1. Neural N/w
2. Economics
3. Image Processing
4. Parallelization
5. Robotics

• Chromosome is also called as individual soln.

Computational Background :



can be a
1. Bit Pattern
: 0/1
2. Real values

Objective Function referred as Fitness Function

Fitness Value

calculated for every chromosome.

Reqd. to get the next set of (population) chromosomes

2 methods :
1. Cross-over
2. Mutation

DATE: PAGE:

1. Population / Set of Chromosome / Set of Individual soln :
Initialization is done. How? Randomly.

2. Compute the fitness value for every individual soln/chromosome.

3. Select the ISE parent chromosome :
One of the following ways can be chosen :
i) Roulette wheel
ii) Tournament Technique
iii) Rank Selection
iv) Random " : Used in Tournament Technique.

4. Cross over : new offsprings

5. ~~Mutation is done :~~ a) Single Point ~~Mutation~~ cross over
b) Multipoint " "
c) Uniform " "

5. Mutation : i) Bit Flop ii) Swap iii) Scramble / Struffle iv) Reverse.

6. Replacement Policy : To replace the new offsprings in the old population to generate next "

Repeat the steps from 2 to 6.
Until the objective is satisfied.

Step 1 : Initialization of Population :
can be done either using bit patterns :

1	0	0	1	0
1	0	1	0	0
⋮				

or
can be done using Real Numbers :

0.3	0.2	0	0	0.3
0.7	0.6	1.2	0.0	
⋮				

DATE: PAGE:

Step 2 : For each of these chromosomes, we calculate the fitness value.

Suppose, the fitness fn is

$ax + by + z = 0$

And population is :

P1	1	2	3
P2	3	4	2
P3	4	9	8
P4	3	2	6
P5	6	9	8
P6	3	2	0

Corresponding fitness value is :

0.25 - (1)

0.29 - (2)

0.3 - (3)

0.23 - (4)

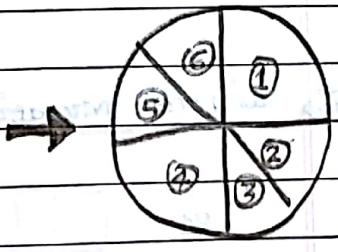
0.05 - (5)

1 - (6)

Step 3 : Choosing Parents :

i) Roulette Wheel :

Rote the wheel twice & whichever comes become the parents.



Suppose Parent 1 : P5

we got 2 " 2 : P1

DATE: PAGE:

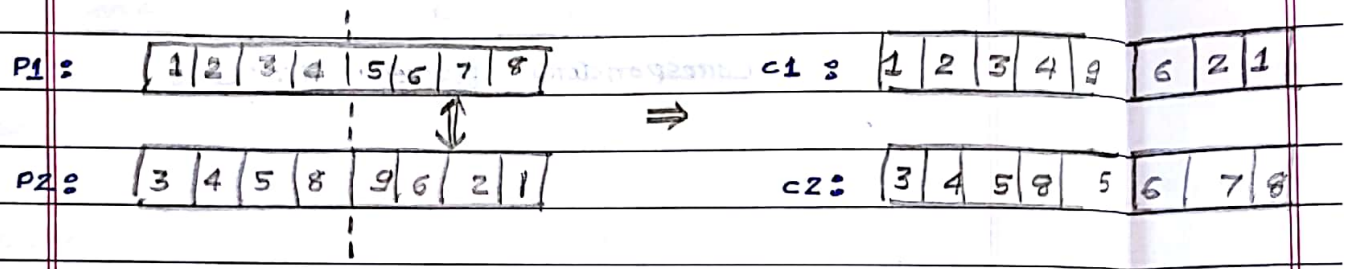
ii) Tournament Technique :

Randomly choose 2 chromosomes.

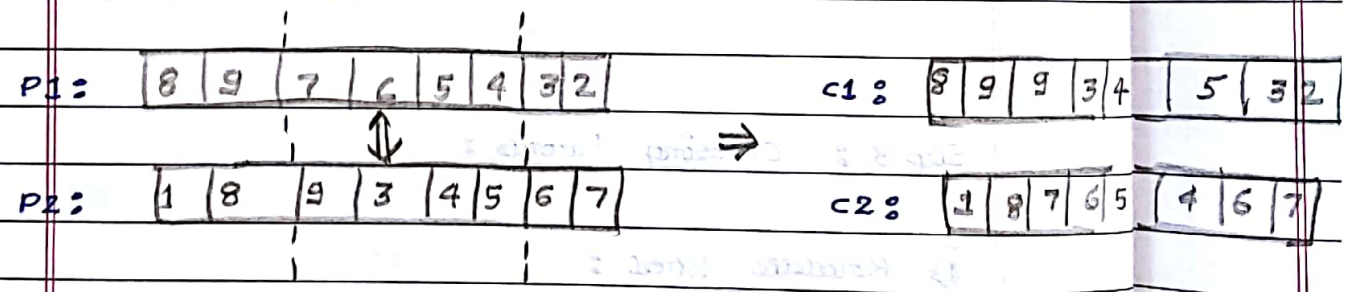
iii) Rank Selection :

Arrange the fitness value is descending order.
Choose chromosomes corresponding to top 2 values as parents.

Step 4 : cross - over : i) Single Point Cross Over



ii) Multipoint Cross Over :



iii) Uniform Mutation : Every gene is treated independently.

P1	0.2	0.2	0.5	0.5	0.3	0.3
P2	0.1	0.8	0.6	0.8	0.7	0.1

iv)

Step

1.

DATE: PAGE:

To get c1's 1st posⁿ,
toss a coin. On tossing, if we get head
fill it with P1s 1st posⁿ
else
P2s 1st posⁿ.

Suppose we tossed a coin 6 times & we got:

H T T H H H

0·2	0·9	0·6	0·5	0·3	0·3

From c1:

c2: will follow

T H H T T T

0·1	0·2	0·5	0·8	0·7	0·1

6 2 1

6 7 8

iv) Out of Syllabus:

$c1 = ax + (1-a)y$

5 3 2

$c2 = ax + (1-a)y$

4 6 7

where, suppose $a = 0.5$

Step 5: Mutation: applied to both the offsprings

1. BIT FLOP:

0	1	0	1	0	1	0	1

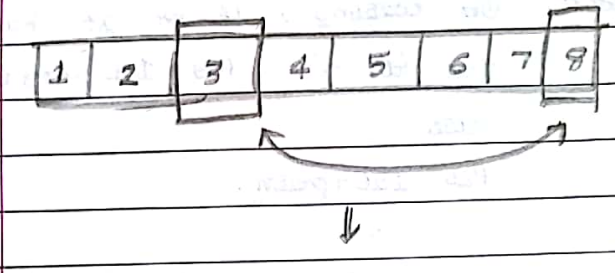
⇓

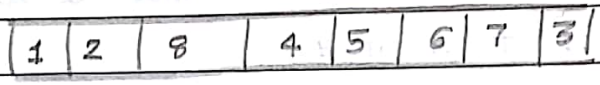
0	1	0	0	0	1	0	1

Posⁿ is chosen randomly.

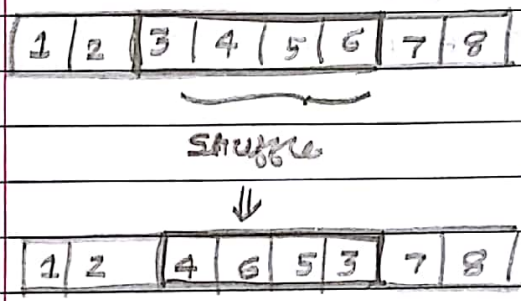
DATE: PAGE:

2. SWAP :

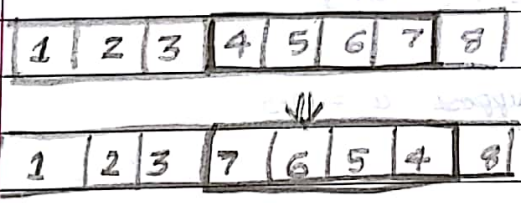




3. Scramble / Shuffle :



4. Reverse :



Step 6 : Replacement Policy

DATE: PAGE:

Termination ways : i) Terminate after fixed # of iterations.

ii) If previous & current population becomes same i.e., there is no improvement anymore, then terminate.

iii) When the set threshold for objective fn is exceeded then terminate.

Numerical Problem :

1. Find 3 nos b/w 1 and 5 (both inclusive) such that $a + 2b + 3c = 15$ where a, b, c are the 3 nos.

Population size = 4

Sol: Initialisation & Fitness Value (Step 1 & 2) :

	a	b	c	$\lvert a+2b+3c-15 \rvert$ = Fitness Value
P1	5	2	3	$\lvert 5+4+9-15 \rvert = 3$
P2	4	3	3	$\lvert 4+6+9-15 \rvert = 4$
P3	2	5	2	$\lvert 2+10+6-15 \rvert = 3$
P4	2	4	5	$\lvert 2+8+15-15 \rvert = 10$

Step 3 : P4 has too high fitness value.

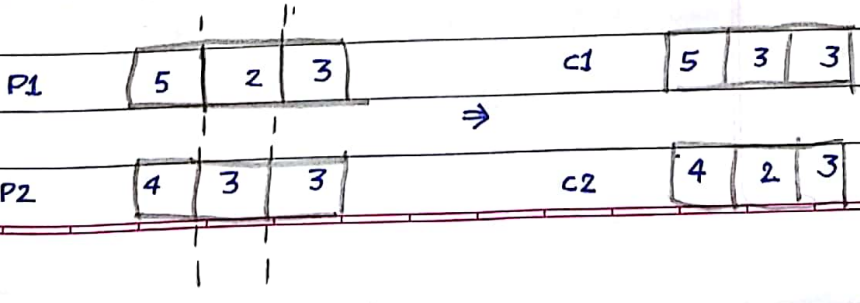
Roulette Wheel b/w P1, P2 & P3

Suppose we get : P1

Now Roulette Wheel b/w P2 & P3

Suppose we get : P2.

Step 4 : Using Single Point crossover b/w P1 & P2



DATE: PAGE:

17th October, 2017

Little diff. from what we discussed

Here, we are finding the new population.

Technique is : i) Keep the parents that made offsprings

&

ii) The offsprings

New population :

P1	5	2	3
P2	4	3	3
c1 : P3	5	3	3
c2 : P4	4	2	3

Now mutate this population

	Fitness Fn	Fitness Value
4 3 3	$\|a+2b+3c-15=0\|$	4+6+9 - 15 = 4
4 3 2		4+6+6 - 15 = 1
5 3 1		5+6+3 -15 = |-1| = 1
3 3 3		3+6+9 -15 = 3

(Done by random replacement mutation technique)

Also with this restart the process

Ex : Total # of

Suppose

Boosting

Ex :

Adaptive Boosting

or

Adaboost

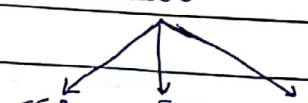
17th October, 2017

Improving Performance of (Binary) Classifier

DATE:

PAGE:

Ex: Total # of students in PESU : 2000



1000 500 500

CS

Suppose 60% of students are recognized by Faculty F1

$= \frac{600}{1000}$

⇓

Classifier 1,

Performance on the entire dataset $= \frac{600}{2000}$

$= 30\%$

30% of EC students are recognized by Faculty F2

$= \frac{150}{500}$

⇓

Classifier 2,

Performance on the entire dataset $= \frac{150}{2000}$

35% of ME students are recognized by Faculty F3

$= \frac{175}{500}$

⇓

Classifier 3,

Performance on the entire dataset $= \frac{175}{2000}$

...ss	Value
+9	$9 \cdot 15 = 4$
+6	$-15 = 1$
+3	$-15 = \lvert -1 \rvert = 1$
+9	$-15 = 3$

Boosting : i) Technique to improve performance on the entire dataset.

ii) It is a technique which improves performance by combination of hypothesis (classifiers) on a sequence

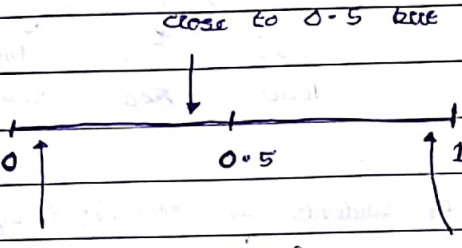
⇓

Boosts the performance

Ex: Adaptive Boosting or Adaboost

DATE: PAGE:

Different types of classifiers / learners based on the Error :

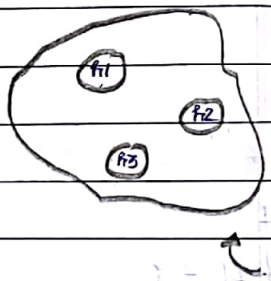


close to zero error : stronger learners

If error is so large, can it be even called a classifier.

Combination of classifiers to improve performance :

If h1, & the part of dataset being wrongly classified by classifier c1, c2, c3 respectively then



h2, h3 are

if we combine result of classifier c1, c2 & c3 & take the votes then, we are likely to get less wrong classification.

can classify h1 & h2 correctly.

can classify h1 & h3 correctly

can classify h2 & h3 correctly

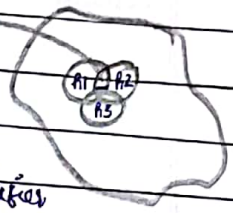
less # of wrong classification means better performance.

DATE: PAGE:

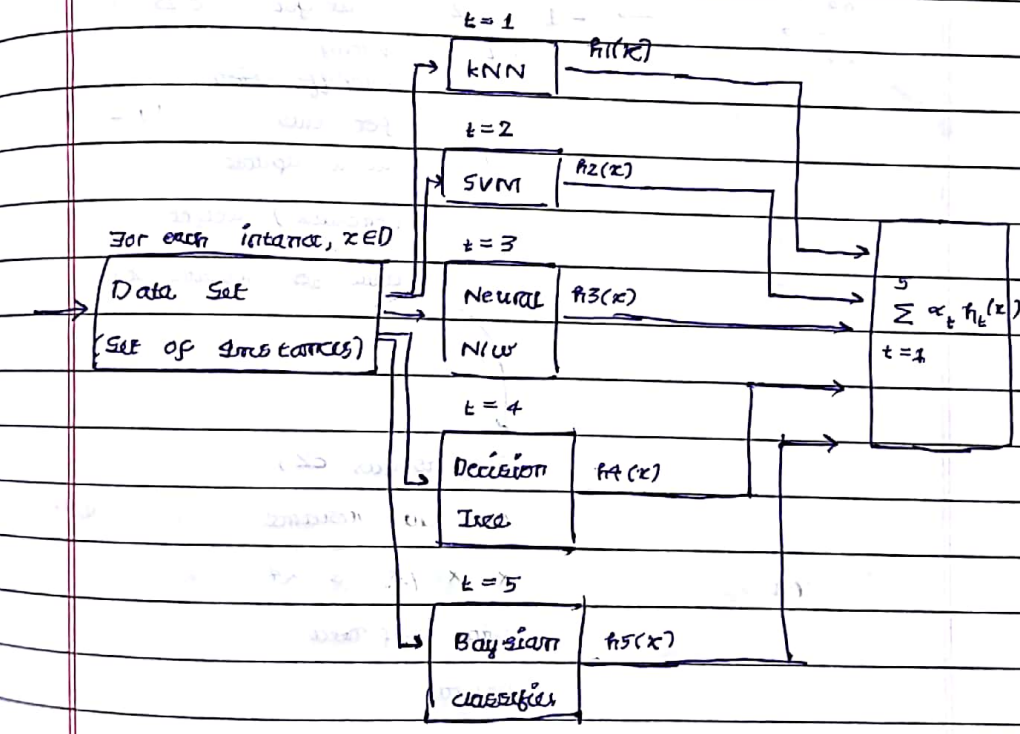
Does combination of classifiers (generally odd # of classifiers) always improve performance ? No

2 out of 3 wrongly classifies. So if you go by majority vote on combn of classifiers in this case will give incorrect classification.

C1 & C2 wrongly classify this region. But C3 classifies classifies this set of examples correctly.



How do we combine classifiers ?



$$H(x) = \alpha_1 h_1(x) + \alpha_2 h_2(x) + \cdots = \sum_{t=1}^{T} \alpha_t h_t(x)$$

where $\alpha = \frac{1}{2} \log_e \left(\frac{1 - E_t}{E_t} \right)$

where E_t is the error

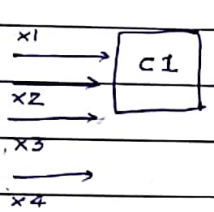
DATE: PAGE:

How it (H(x)) is actually calculated ?

Consider	True label (y_i)
x1	+1
x2	-1
x3	-1
x4	+1

Initial values of $w = \frac{1}{\#\text{ of instances}} = \frac{1}{4} = 0.25$

Considering classifier c1



$h(x_i)$	$\hat{y}_i$	Error	weight values (w)
→	+1	0	0.25
→	-1	0	0.25
→	-1	0	0.25
→	-1	2	0.25

∴ we got wrong classification for this we'll update (increase) weight value for instance x4

To classify c2, we feed instance x1, x2, x3 & x4 with Updated weights

NOTE : w value is associated with instance values.

DATE: PAGE:

Ex: Out of 10 samples,

6 are correct

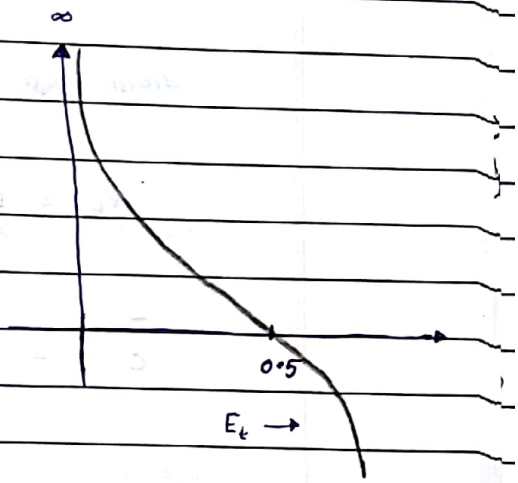
4 are wrong

Thm: $E = \frac{4}{10} = 0.4$

$$\alpha = \frac{1}{2} \log_e \left(\frac{1-0.4}{0.4} \right) = \frac{1}{2} \log_e \left(\frac{0.6}{0.4} \right).$$

Plotting E_t vs α: $E_t \in [0, 1]$

$$\alpha_t = \frac{1}{2} \log_e \left(\frac{1 - E_t}{E_t} \right) \quad -(2)$$



If $E_t = 0$,

$$\alpha = \frac{1}{2} \log_e \left(\frac{1}{0} \right) = \infty$$

If $E_t = 1/2$,

$$\alpha = \frac{1}{2} \log_e \left(\frac{1 - 1/2}{1/2} \right) = 0$$

If $E_t > 0.5$,

$$\alpha = \frac{1}{2} \left[\log_e (\text{small}) - \log_e (\text{large}) \right] < 0$$

DATE: PAGE:

Derivation :

$$w_{t+1}^{i} = \frac{w_t^{i}}{Z} \times e^{-\alpha_t h_t(i) y_i}$$

Substituting α in abv eqn,

$$w_{t+1}^{i} = \frac{w_t^{i}}{Z} \times e^{-1/2 \log_e \left(\frac{1-E_t}{E_t}\right) h_t(i) y_i}$$

$$\Rightarrow w_{t+1}^{i} = \frac{w_t^{i}}{Z} e^{-1/2}$$

Note :

$h_t(i)\, y(i)$

$\equiv \hat{y}(i)\, y(i)$

for correct classification is always equal to $+1$

From eqn (2),

$$\alpha_t = \frac{1}{2} \log_e \left(\frac{1-E_t}{E_t}\right)$$

$$e^{-\alpha_t} = e^{-1/2 \log_e \left(\frac{1-E_t}{E_t}\right)} = \sqrt{\frac{1-E_t}{E_t}}$$

For wrong classification ,

$$e^{-\alpha_t} = \left(\frac{1-E_t}{E_t}\right)^{1/2}$$

$$Z = (1-E_t)\sqrt{\frac{E_t}{(1-E_t)}} + E_t\sqrt{\frac{1-E_t}{E_t}}$$

$$= (1-E_t)\sqrt{\frac{E_t}{(1-E_t)}} + E_t\sqrt{\frac{(1-E_t)}{E_t}}$$

$$= (\sqrt{1-E_t})(\sqrt{E_t}) + (\sqrt{E_t})(\sqrt{1-E_t})$$

$$= 2(\sqrt{E_t})(\sqrt{1-E_t})$$

$$Z = 2\sqrt{E_t(1-E_t)}$$

Similarly, $\sum w_{t+1}^{i} = (1$ wrong samples

Simplifying we get : $2\sqrt{E_t(1-E_t)} \cdot \sum_{\text{correct samples}} w_{t+1}^i = \sum_{t=1}^{T} w_t^i \sqrt{\frac{E_t}{1-E_t}}$

DATE: PAGE: 1/2

$$\sum_{t=1}^{T} w_{t+2}^i = \sum_{t=1}^{T} \frac{w_t^i}{Z} \sqrt{\frac{E_t}{1-E_t}} \quad - (a)$$ for all correct instances

$$\sum_{t=1}^{T} w_{t+1}^i = \sum_{t=1}^{T} \frac{w_t^i}{Z} \sqrt{\frac{1-E_t}{E_t}} \quad - (b)$$ for all wrong instances

similarly $\sum_{\text{wrong samples}} w_{t+1}^i = 1/2$

(a) + (b)

$$Z \left\{ \sum_{i \text{ correct instances}} w_{t+1}^i + \sum_{i \text{ wrong instances}} w_{t+1}^i \right\}$$

$$= \left\{ \sum_{t=1}^{T} w_t^i \sqrt{\frac{E_t}{(1-E_t)}} \right\}$$ correct instances

$$+ \left\{ \sum_{t=1}^{T} w_t^i \sqrt{\frac{(1-E_t)}{E_t}} \right\}$$ wrong instances

DATE: PAGE:

Example : Consider the classification training as follows :

x →	0	1	2	3	4	5	6	7	8	9
y: classification →	+1	+1	+1	-1	-1	-1	+1	+1	+1	-1

Check the accuracy against single hypothesis (classifier)

Use 3 classifiers,

$$H_t(x_i) = \text{sign}\left\{ \sum_{t=1}^{T} \alpha_t h_t(x_i) \right\}$$

# of instances $N = 10$

Sol: Since we have 3 classifiers,

$T = 3$

So,

$$H_t(x_i) = \text{sign}\left\{ \alpha_1 h_1(x_i) + \alpha_2 h_2(x_i) + \alpha_3 h_3(x_i) \right\}$$

$$\alpha = \frac{1}{2} \log_e \left(\frac{1 - E_t}{E_t} \right)$$

Assuming classifier 1, does the following classification :

x →	0	1	2	3	4	5	6	7	8	9
$\hat{y}$ →	+1	+1	+1	-1	-1	-1	-1	-1	-1	-1

split is at $\left(\frac{2+3}{2}\right) = 2.5$ i.e., +ve samples are to left of 2.5, -ve " " " right

initial weights associated with each of these instance,

$$w = \frac{1}{N} = \frac{1}{10} = 0.1$$

x_i →	0	1	2	3	4	5	6	7	8	9
w_i →	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
$\hat{y}_i$ →	+1	+1	+1	-	-	-	-	-	-	-
y_i →	+	+	+	-	-	-	+	+	+	-

→ wrong classification by classifier 1

∵ we have 3

∴ E_t = 0.

$1 - E_t$ =

So α value

α_1 =

=

=

∴ 1st term

H_t

Now updating

$w_{updated}$ $\sum w_i$

for +ve for all x_i with correct

$w_{updated}$ [0.1 ×

for +ve

$w_{updated}$

w f

∵ We have 3 wrong classification ; DATE: PAGE:

∴ $E_t = 0.1 + 0.1 + 0.1 = 0.3$

$1 - E_t = 1 - 0.3 = 0.7$

So α value for classifier 1 ,

$$\alpha_1 = \frac{1}{2} \log_e \left(\frac{1 - E_t}{E_t} \right)$$

$$= \frac{1}{2} \log_e \left(\frac{0.7}{0.3} \right)$$

$$= 0.424$$

∴ 1st term of

$H_t(x_i) \Rightarrow \alpha_1 h_1(x_i)$

becomes

$\boxed{0.424\, h_1(x_i)}$

Now updating (decreasing w_i values in case classifier 1 c1 made correct classification & increasing w_i values in case classifier 1 c1 made correct classification) ,

$w_{updated}$ for +ve for all x_i with correct classification $\times \sum w_i \cdot$ = 1/2

$w_{updated}$ for +ve $\times [0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1] = 1/2$

$w_{updated} = \frac{1}{2 \times 0.7} = 0.714$

w for +ve samples of classifier 1 , c1 = 0.1×0.714

$= 0.0714$ -(1)

2.5 of 2.5

ation by classifier 1

For wrong classification,

DATE: PAGE:

$w' \Sigma w_i = 1/2$ for all x_i for which classifier 1 wrongly classified

$w' [0.1 + 0.1 + 0.1] = 1/2$

$w' \times 0.1 \times 3 = 1/2$

$w' = \frac{1}{2 \times 3 \times 0.1} = \frac{1}{0.6} = \frac{10}{6} = 1.666$

updated w values for the wrong classification done by classifier 1,

0.1×1.666

$= 0.0166$ - (2)

Assuming classification done by 2nd classifier is

split at 8.5

x_i	0	1	2	3	4	5	6	7	8	9
$\hat{y}_i$	+	+	+	+	+	+	+	+	+	-
y_i	+	+	+	-	-	-	+	+	+	-

# of wrong classification made by classifier 2 = 3

Using (1) & (2),

w_i	0.07	0.07	0.07	0.07	0.07	0.07	0.166	0.166	0.166	0.07

∵ we have 3 wrong classification,

$E_t = 0.07 + 0.07 + 0.07$

$= 0.21$

$1 - E_t = 1 - 0.21 = 0.79$

So α value for classifier 2,

$\alpha_2 = \frac{1}{2} \log_e \left(\frac{1 - E_t}{E_t} \right) = \frac{1}{2} \log_e \left(\frac{0.79}{0.21} \right) = 0.653$

DATE: PAGE:

∴ 2 term of $H_L(x_i)$ is $\alpha_2 h_2(x_i)$

$= 0.663\, h_2(x_i)$

On the next classification, the initial values of w needs to be updated.

Before next classification happens

~~w value for x1, x2, x3, x6, x7, x8 & x9~~

should be $\frac{1}{2} \times \frac{1}{(0.071 \times 4) + (0.166 \times 3)}$

$= \frac{1}{2\,[0.284 + 0.498]}$

$= \frac{1}{2 \times 0.782}$

$= 0.639$

$\simeq 0.64$

w for x0, x1, x2, ~~x6, x7, x8, x9~~ $= 0.071 \times 0.64$

x9 $= 0.045$

w for x3, x4, x5 $= 0.071 \times 2.34$

$= 0.16614$

w for x6, x7, x8, x9 $= 0.166 \times 0.64$

$= 0.106$

For classification by C3, (suppose split at 5.5)

x_i	0	1	2	3	4	5	6	7	8	9
y_i	+	+	+	+	+	+	-	-	-	-
y_i	+	+	+	-	-	-	+	+	+	-
w_i	0.045	0.045	0.045	0.166	0.166	0.166	0.106	0.106	0.106	0.045

DATE: PAGE:

Now # of correct classification = 4

# of incorrect " = 6

E_t = 0.166 + 0.166 + 0.166 + 0.166 + 0.166 + 0.166

= 0.498 + 0.318

= 0.816

$1 - E_t = 0.184$

So α value for classifier 3,

$$\alpha_3 = \frac{1}{2} \log_e \left(\frac{1 - E_t}{E_t} \right)$$

$$= \frac{1}{2} \times \log_e \left(\frac{0.184}{0.816} \right)$$

$$= -0.744$$

$\therefore H_t$

$$= \text{sign} \{ \alpha_1 h_1(x_i) + \alpha_2 h_2(x_i) + \alpha_3 h_3(x_i) \}$$

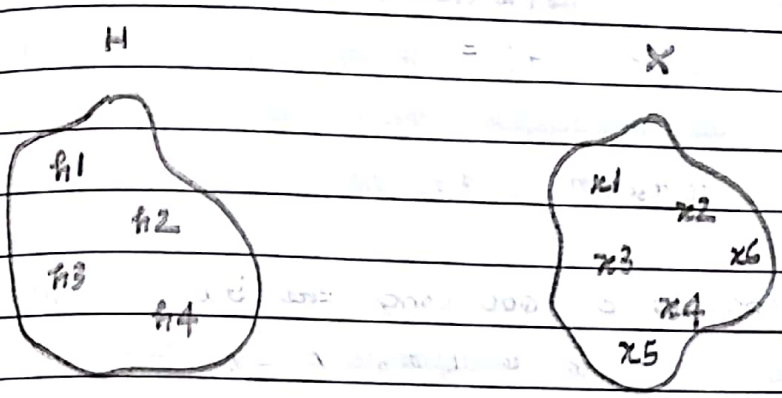
$$= \text{sign} \{ 0.424 \, h_1(x_i) + 0.663 \, h_2(x_i) - 0.744 \, h_3(x_i) \}$$

=

Bayesian Learning

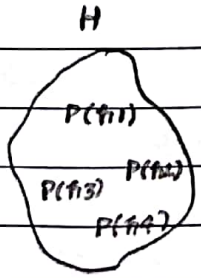
DATE: PAGE:

How is Bayesian Learning different from something like concept learning algs such as List then Eliminate or Candidate Elimination Algs ?



In algorithms such as,
List then Eliminate or Candidate Elimination,
from hypothesis set H we try to eliminate all $h \in H$ that is not consistent with all $x \in X$. The set of hypothesis left in H is referred to as Version Space.

In Bayesian learning,
instead of eliminating $h \in H$ which are not consistent with all $x \in X$,
i) we consider all $h \in H$
ii) initialize each one them with equal probability
iii) The req. Version Space will consist of all the hypothesis $h \in H$ with max. probabilities



DATE: PAGE:

Computation of Probabilities :

i) For all $h \in H$,
initially $P(h) = \frac{1}{\|H\|}$
where $\|H\|$ = # of hypothesis in H
So, all hypothesis $h \in H$ are initialized with uniform probabilities.

ii) Probability of observing the data, $P(D)$
Thus "" is independent of the hypothesis

iii) Prior probability, $P(D|h)$
is the "" that h has been observed for data D

To find our version space, we will calculate Probability of h given data set D for each $h \in H$ & then find the hypothesis (or set of hypothesis) having max. probability.

Brute Force Method : This method assumes that the dataset D used for training was/is COMPLETLY ERROR FREE

Comp-uta-tionally expensive

Posterior Probability of h given dataset D, $P(h|D)$

$$= \frac{P(D|h)\,P(h)}{P(D)}$$

{ using Baye's Theorem }

we want to find,

Maximum aposteriori hypothesis in the term used when we assumed that our training data was completely error free,

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$

$$= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

$\because P(D)$ is constant, we take it off.

DATE: PAGE:

Numerical on Bayesian Learning :

Q. 0.008 of the population is suffering from cancer. A lab conducts tests on patients to conform this. When the patient has cancer, 98% of the time the lab results are +ve, & when the patient doesn't have cancer 97% of the time the lab results are -ve. What is the probability that the patient has cancer if the lab result is +ve ?

Sol. $P(\text{patient has cancer}) = 0.008$

$P(\text{lab result is +ve} \mid \text{patient has cancer}) = 0.98$

$P(\text{lab result is -ve} \mid \text{patient does not have cancer}) = 0.97$

$P(\text{patient has cancer} \mid \text{lab result is +ve})$

$$= \frac{P(\text{lab result is +ve} \mid \text{patient has cancer}) \times P(\text{patient has cancer})}{P(\text{lab result is +ve})}$$

Observation : $P(h|D)$ for any $h \in H$

$$\propto P(D|h)$$
$$\propto P(h)$$
$$\propto \frac{1}{P(D)}$$

$P(\text{lab result is +ve})$

$= P(\text{person has cancer})\,P(\text{lab result is +ve} \mid \text{person has cancer}) + P(\text{person does not have cancer})\,P(\text{lab result is +ve} \mid \text{person does not have cancer})$

DATE: PAGE:

$= 0.008\ (0.98)$

$+$

$(1-0.008)(1-0.97)$

$= 0.008(0.98) + (0.992)(0.03)$

$= 0.00784 + 0.02976$

$= 0.0376$

$\therefore P(\text{patient has cancer} \mid \text{lab result is +ve})$

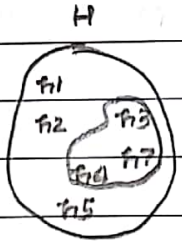
$= \dfrac{0.98\ (0.008)}{0.0376}$

$= \dfrac{0.00784}{0.0376}$

$= 0.2085$

— x — x — x —

Derivation :



Suppose Version Space VS contains h3, h4 & h7 then

$P(D|h) = 1$

for all $h \in VS$.

&

$P(D|h) = 0$

for all $h \notin VS$

WKT,

DATE: PAGE:

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

Initially, $P(h) = \frac{1}{\|H\|}$

$$\therefore P(h|D) = \frac{P(D|h) \times \frac{1}{\|H\|}}{P(D)}$$

$$= \frac{P(D|h) \times \frac{1}{\|H\|}}{\sum_{h \in VS} P(D|h)P(h) + \sum_{h \notin VS} P(D|h)P(h)}$$

$$= \frac{P(D|h) \times \frac{1}{\|H\|}}{\sum_{h \in VS} 1 \cdot \frac{1}{\|H\|} + \sum_{h \notin VS} 0 \cdot \frac{1}{\|H\|}}$$

$$= \frac{P(D|h) \times \frac{1}{\|H\|}}{\frac{1}{\|H\|} \sum_{h \in VS} 1 + 0}$$

$$= \frac{P(D|h) \times \frac{1}{\|H\|}}{\frac{\|VS\|}{\|H\|}}$$

$$= \frac{1}{\|VS\|}$$

{ $\because P(D|H) = 1$ if consistent }

Bayesian Rule (continued)

DATE: PAGE:

Till now we calculated h_{MAP} under the assumption that the training data that we had was completely error free.

Now we would calculate h_{ML}, knowing that the training data contains error.

In case our training data is

x_i	Attr 1	Attr 2 ...	True label (d_i & not y_i)
x_0			d_0
x_1			d_1
x_2			d_2
:			:
x_n			d_n

Error could either be in i) the attribute value of a particular instance.

ii) true label of a particular instances.

While calculating h_{ML} we work under the assumption that few instances of the training data may have wrong true labels, i.e., for few examples in the training data,

$$d_i = F(x) + e_i$$

d_i: True label which should have been there

DATE: PAGE:

Lagrange's Theorem:

i) Used for minimizing / maximizing a multivalued fn $f(x, y)$ subjected to constraint $g(x, y)$

ii) This theorem states that,

gradient vector of $f(x, y)$ is parallel to the gradient vector of $g(x, y)$

iii) Lagrange's fn,

$L(x, y) = f(x, y) - \alpha g(x, y)$ where $\alpha \geq 0$

and

$\nabla f(x, y) = \alpha \nabla g(x, y)$

$L(x, y, \alpha)$ is expressed as $\theta(\alpha)$

where x & y are expressed in terms of α.

Also,

$\nabla L = 0$

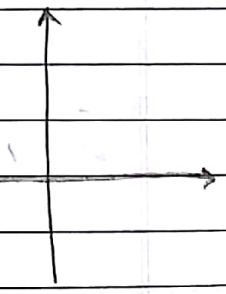
$\Rightarrow \nabla f - \alpha \nabla g = 0$ where $\alpha \geq 0$

Example: Minimize $f(x, y) = x^2 + y^2$

subject to

constraint

$g(x, y) \equiv y + x - 1 = 0$



Sol. $L(x, y) = f(x, y) - \alpha g(x, y)$

where $\alpha \geq 0$

&

$\nabla L = \nabla f - \alpha \nabla g = 0$

So, $L(x, y) = (x^2 + y^2) - \alpha(y + x - 1)$

DATE: PAGE:

$$\nabla f = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

$$\nabla g = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

By Lagrange,

$$\nabla L = 0; \quad \alpha \geq 0$$

$$\Rightarrow \quad \nabla f - \alpha \nabla g = 0$$

$$\Rightarrow \quad \begin{bmatrix} 2x \\ 2y \end{bmatrix} - \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 0$$

$$\Rightarrow \quad 2x - \alpha = 0 \quad \& \quad 2y - \alpha = 0$$

So, $\boxed{x = \frac{\alpha}{2}}$ So, $\boxed{y = \frac{\alpha}{2}}$

Expressing $L(x, y) = f(x, y) - \alpha g(x, y)$

where $\alpha \geq 0$

as

$\Theta(\alpha)$.

So, $L(x, y) = (x^2 + y^2) - \alpha(y + x - 1)$

$$\Theta(\alpha) = \left(\frac{\alpha^2}{4} + \frac{\alpha^2}{4}\right) - \alpha\left(\frac{\alpha}{2} + \frac{\alpha}{2} - 1\right)$$

$$= \frac{\alpha^2}{4} + \frac{\alpha^2}{4} - \frac{\alpha^2}{2} - \frac{\alpha^2}{2} + \alpha$$

$$= \frac{\alpha^2}{2} - \frac{\alpha^2}{2} - \frac{\alpha^2}{2} + \alpha$$

$$= \alpha - \frac{\alpha^2}{2}$$

DATE: PAGE:

According to Lagrange,

maximizing $\Theta(\alpha)$

is

same as

minimizing $f(x, y)$

Minimizing $\Theta(\alpha)$,

$$\Theta'(\alpha) = 0$$

$$1 - \frac{2\alpha}{2} = 0$$

$$\boxed{\alpha = 1}$$

At $\left(x = \frac{\alpha}{2}, y = \frac{\alpha}{2}\right)$, $f(x, y)$ is minimum

$= \left(\frac{1}{2}, \frac{1}{2}\right)$ $f(x, y)$ is minimum.

subjected to constraint

$g(x, y)$.

Ex 2: $f(x, y) = x^2 y$ subjected to constraint $g(x, y)$:

$$x^2 + y^2 - 1 = 0$$

Sol. $L(x, y) = f(x, y) - \alpha g(x, y)$ where $\alpha \geq 0$

&

$$\nabla L = 0$$

$$\Rightarrow \nabla f - \alpha \nabla g = 0 \quad \Rightarrow \quad \begin{bmatrix} 2xy \\ x^2 \end{bmatrix} - \alpha \begin{bmatrix} 2x \\ 2y \end{bmatrix} = 0$$

So, $2x[y - \alpha] = 0$

$x = 0$ or $\boxed{y = \alpha}$

$$x^2 - 2\alpha y = 0$$

$$x^2 = 2\alpha y$$

$$x^2 = 2\alpha^2$$

$$\boxed{x = \pm\alpha\sqrt{2}}$$

$\left(\frac{x}{\alpha}\right)^2 =$

DATE: PAGE:

$\theta(\alpha) = 2\alpha^2(\alpha) - \alpha[2\alpha^2 + \alpha^2 - 1]$

$= 2\alpha^3 - \alpha[3\alpha^2 - 1]$

$= 2\alpha^3 - 3\alpha^3 + \alpha$

$\theta(\alpha) = \alpha - \alpha^3$

Minimizing $f(x,y)$ wrt $g(x,y)$ is same as maximizing $\theta(\alpha)$

$\theta'(\alpha) \Rightarrow 1 - 3\alpha^2 = 0$

$\Rightarrow 3\alpha^2 = 1$

$\Rightarrow \alpha^2 = \frac{1}{3}$

$\Rightarrow \alpha = \pm \frac{1}{\sqrt{3}}$

But $\alpha \geq 0$

$\therefore \alpha = \frac{1}{\sqrt{3}}$

$\therefore$ At $(2\alpha^2, \alpha)$ $f(x,y)$ is min. subjected to constraint $g(x,y)$

$\left(\frac{2}{3}, \frac{1}{\sqrt{3}}\right)$ $f(x,y)$ is min subjected to constraint $g(x,y)$

Bayesian learning DATE: PAGE:

(Contd)

Given : $P(h_1|D) = 0.4$, given its +ve.

$P(h_2|D) = 0.3$, " " its -ve.

$P(h_3|D) = 0.3$, " " " -ve.

Suppose a new instance x_i is given.

If we use h_{MAP}, then

classification of this sample is +ve

$\because P(h_1|D)$ is max.

If we use "Bayes Optimal classifier" then, $V = \{+, -\}$

$$P(V_j|D) = \sum_{h_i \in H} P(V_j|h_i) \times P(h_i|D)$$

So, $P(+|D) = P(+|h_1)P(h_1|D) + P(+|h_2)P(h_2|D) + P(+|h_3)P(h_3|D)$

$= 1(0.4) + 0 + 0$

$= 0.4$

$P(+|D) = 0.4$

$P(-|D) = P(-|h_1)P(h_1|D) + P(-|h_2)P(h_2|D) + P(-|h_3)P(h_3|D)$

$= 0(0.4) + 0.3(1) + 0.3(1)$

$P(-|D) = 0.6$

$P(+|D) = 0.4$

$P(-|D) = 0.6$ } max is -ve

$\therefore$ Classification is -ve.

Bayes Optimal classifier is better than h_{MAP}.

DATE: PAGE:

Gibbs Algorithm:

1. Choose one hypothesis 'h' at random.
2. The classification of the new instance is wrt the hypothesis 'h' in Step 1.

Error $\left(\frac{\text{\# of wrong classification}}{\text{Total \# of ''}}\right)$ because of Gibbs Algorithm cannot be more than 2 times Error in Bayes Optimal Classifier Alg.