Big Data Exam

Q1

A)

TID	Items
T1	K A D B
T2	D A C E B
T3	C A D B
T4	A B D
T5	C D E

Initially, $C_1 = \{A, B, C, D, \cancel{E}, \cancel{K}\}$

$(4)\ (4)\ (3)\ (5)\ (2)\ (1)$

Min Supp = 3

$>3\ >3\ =3\ >3\ <3\ <3$

$L_1 = \{A, B, C, D\}$

$C_2 = L_1 \bowtie L_1$ of size 2

$= \{AB, \cancel{AC}, AD, \cancel{BC}, BD, CD\}$

$(4)\ (2)\ (4)\ (2)\ (4)\ (3)$

$>3\ <3\ >3\ <3\ >3\ =3$

$L_2 = \{AB, AD, BD, CD\}$

$C_3 = L_2 \bowtie L_2$ of size 3

$= \{ABD, \cancel{ACD}, \cancel{BCD}\}$

$(4)\ (2)\ (2)$

$>3\ <3\ <3$

$L3 = \{ ABD \}$

~~Stop~~.

$C4 = L3 \bowtie L3$ of size 4

$= \{ \}$ empty.

Stop.

Freq Itemsets $= L1 \cup L2 \cup L3$

$= \{ A, B, C, D, AB, AD, BD, CD, ABD \}$

(B)

(B) Given conditions, For a rule $X \rightarrow Y$,

$\rightarrow$ Support $= \dfrac{\text{Count}(X \cup Y)}{N} \in [0.3, 0.5]$

$\rightarrow$ Accuracy $= \dfrac{\text{Count}(X \cup Y)}{\text{Count}(X)} > 0.6$

So, we can apply a refined Apriori algorith. where we accept items from C_i to L_i only if $\dfrac{\text{Supp}(\text{Itemset})}{N} \in [0.3, 0.5]$

and after finding such itemsets, we define rules $X \rightarrow Y$ using those itemsets such that their confidence > 0.6

Applying this on the dataset,

Initially, $C_1 = \{A, B, C, D, E, K\}$

Here, $N = 5$, so count $\in [0.3 \times 5, 0.5 \times 5]$

$\Rightarrow$ count $\in [1.5, 2.5]$

as count is a integer $\Rightarrow$ count must be $= 2$.

$$\therefore C_1 = \{ \cancel{A}, \cancel{B}, \cancel{C}, \cancel{D}, E, \cancel{K} \}$$
$$\quad (4) \quad (4) \quad (3) \quad (5) \quad (2) \quad (1)$$
$$\quad >2 \quad >2 \quad >2 \quad >2 \quad =2 \quad <2$$

$L_1 = \{E\}$

$C_2 = L_1 \bowtie L_1$ of size 2

$= \{\}$ empty.

So, Stop.

$\therefore$ Accepted Itemsets $= \{E\}$

$\therefore$ Initially we generate all rules,

$\Rightarrow$ ① $\{\} \rightarrow E$

$$\text{confidence} = \frac{\text{Supp}(E)}{\text{Supp}(\{\})} = \frac{2}{5} = 0.4$$

as $0.4 < 0.6$, this rule is not accepted.

② $E \rightarrow \{\}$

$$\text{confidence} = \frac{\text{Supp}(E)}{\text{supp}(E)} = 1$$

as $1 > 0.6$, this rule is accepted.

1. c)

i) For FpGrowth,

Time Complexity = $O(n^2)$

where n = no. of unique items in dataset

as in Fpgrowth, we search paths in FpTree for each element in Header Table

$\Rightarrow$ No. of elements in Header table = $O(n)$

Max Tree Depth = $O(n)$

$\therefore \Rightarrow O(n) \cdot O(n) = O(n^2)$

Space Complexity = $O(n^2)$

as we need to construct FpTree which at worst case contains $O(n^2)$ nodes and thus that is space comp where n is no. of unique items in dataset.

ii) For A Close,

As it is similar to Apriori algorithm,

Time Complexity = $O(2^n)$

where n = no. of unique items in dataset

as in worst case when all combinations are frequent, we need to check support for 2^n itemsets.

Space Complexity = $O(2^n)$

as we need to store all C_i and L_i.

and in worst case elements in C_1, C_2 are

$^nC_1$, $^nC_2$, ...

So, total space = $^nC_1 + \dots {}^nC_n = O(2^n)$.

tasset

n

Table.

So, rules => only $E \rightarrow \{\}$

D) <u>Statement:-</u>

Downward Closure Property of Apriori algorithm states that if an itemset I is frequent, then all of it's subsets are also frequent.

<u>Proof:-</u>

Proof :-

Consider a freq itemset I.

Now, consider any subset of I, A.

$\therefore A \subseteq I$.

we can also write, $A \cup \bar{A} = I$

as I is freq,

Supp Count $(I) \geq$ minSupp. —— ①

i.e. Count $(A \cup \bar{A}) \geq$ minSupp

From ①,

Count of all items in I <u>appearing together</u> $\geq$ minSupp

$\therefore$ Every combination of items in I appears atleast Count(I) as every combination of items are included in I.

i.e. Suppose $I = i_1 i_2 \ldots i_K$

and count $(I) = N$

Then any combination of $\{i_1, i_2, \ldots i_K\}$ appear atleast N times as they appear as part of I. (If ABC occurs 3 times, then AB occurs atleast 3 times)

eg. $i_1 i_2 i_3$ is part of $i_1 \ldots i_K$

and so appears atleast N times. $\geq$ minSup

As all subsets are a part of I,

$\therefore$ A appears atleast Count(I) times $\geq$ minSupp

So, A is also freq.

$\therefore$ Every Subset of freq itemset is freq.

2.

A) $$P(H \mid X) = \frac{P(H)\,P(X \mid H)}{P(X)}$$

Here, X - Input data

H - Target / Prediction

This equation can be simplified as,

$P(H \mid X)$ = Probability that required target is H given some input data X.

$P(H)$ = Likelihood of appearance of H as target value

$P(X)$ = Likelihood of appearance of X as input data

$P(X \mid H)$ = Probability that input data is X given that target value = H.

In classification, training data has 2 parts input features X and their target H.

eg.

x_1	x_2	H
1	2	3
4	5	6

And test data contains only input data and we need to predict H value using input X.

In training phase,

as we know both X and H values,

∴ Using training data,

we know $P(H|X)$, $P(H)$ and $P(X)$.

we estimate $P(X|H)$ for all categories and values of H.

Then in testing phase,

we find $P(H|X)$ for each category of H using the formula as we know other parameters from training phase.

we predict target as H with maximum $P(H|X)$.

So, $P(H|X)$ = Prediction

$P(H)$ = Likelihood of target

$P(X)$ = evidence of input

$P(X/H)$ = Prior Knowledge

B) It is called 'Naive' Bayes Classifier as we assume (naively) that all attributes of data points are mutually independent.

i.e. For attributes $x_1, x_2, \dots x_n$,

$$x_i \cap x_j = \phi \quad \forall\ i, j \in N \text{ such that } 1 \le i, j \le n,\ i \neq j$$

Graphical Model,

Class - C

Attr - A_1, A_2, A_3, A_4 (mutually independent)

2. c) From figure,

y = binary $\rightarrow$ {Yes, No}

$\therefore P(y = Yes) = \frac{5}{9}$

$P(y = No) = \frac{4}{9}$

Then we find $P(x_i / y)$ for each column.

We convert Age column into categories

{ (20-30), (30-40), (40-50) }

So,

Age	Loan Yes	No	P(Yes)	P(No)
20-30	1	1	1/5	1/4
30-40	2	1	2/5	1/4
40-50	2	2	2/5	2/4
Total	5	4		

Income	Loan Yes	No	P(Yes)	P(No)
Low	1	3	1/5	3/4
Med	3	0	3/5	0
High	1	1	1/5	1/4
Tot	5	4		

Marital	Loan Yes	No	P(Yes)	P(No)
Yes	3	2	3/5	2/4
No	2	2	2/5	2/4
Tot	5	4		

Cred	Loan Yes	No	P(Yes)	P(No)
Fair	3	2	3/5	2/4
Exc	2	2	2/5	2/4
Tot	5	4		

Now, for test example.

(35, Medium, Yes, Fair)

→ (30-40, Medium, Yes, Fair)

$P(Yes|x) \propto P(Yes) \cdot P(30\text{-}40/Yes)\, P(med/Yes)\, P(Yes/Yes)\, P(Fair/Yes)$

$\propto \frac{5}{9} \cdot \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{3}{5}$

$\propto \frac{270}{5625} \propto 0.048$

$P(No|X) \propto P(No) \cdot P(30\text{-}40/No)\, P(med/No)\, P(Yes/No)\, P(Fair/No)$

$\propto \frac{4}{9} \cdot \frac{1}{4} \cdot 0$

$\propto 0$.

Applying Laplace Correction,

$P(med/No) = 1/4$

$\therefore P(No/X) \propto \frac{4}{9} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{2}{4} \propto \frac{16}{2304} \propto 0.006$

$\therefore$ Clearly $P(\text{Yes} \mid X) > P(\text{No} \mid X)$

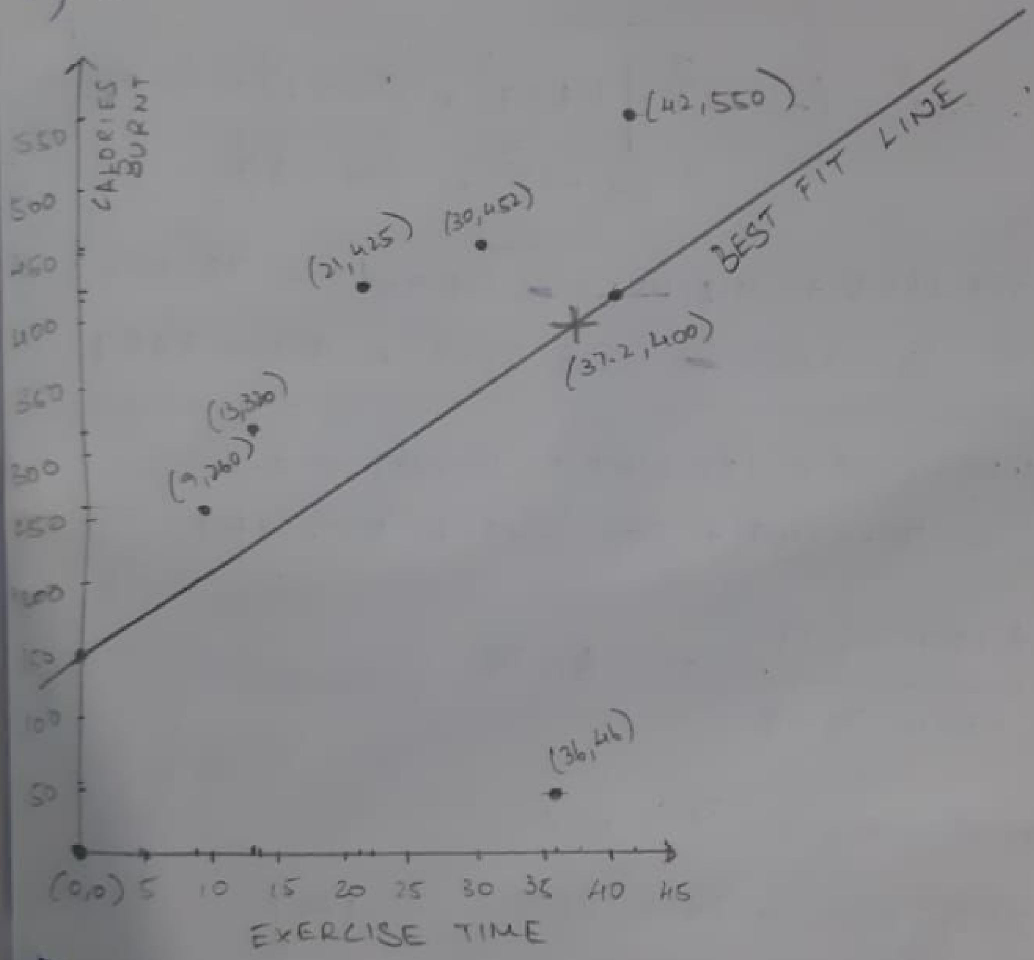
$\therefore$ Prediction = Yes

8)

4. a)

Points are, (0,0), (9,260), (13, 320), (21,425)
(30,452), (36,46), (42,550)

i) Scatter Plot



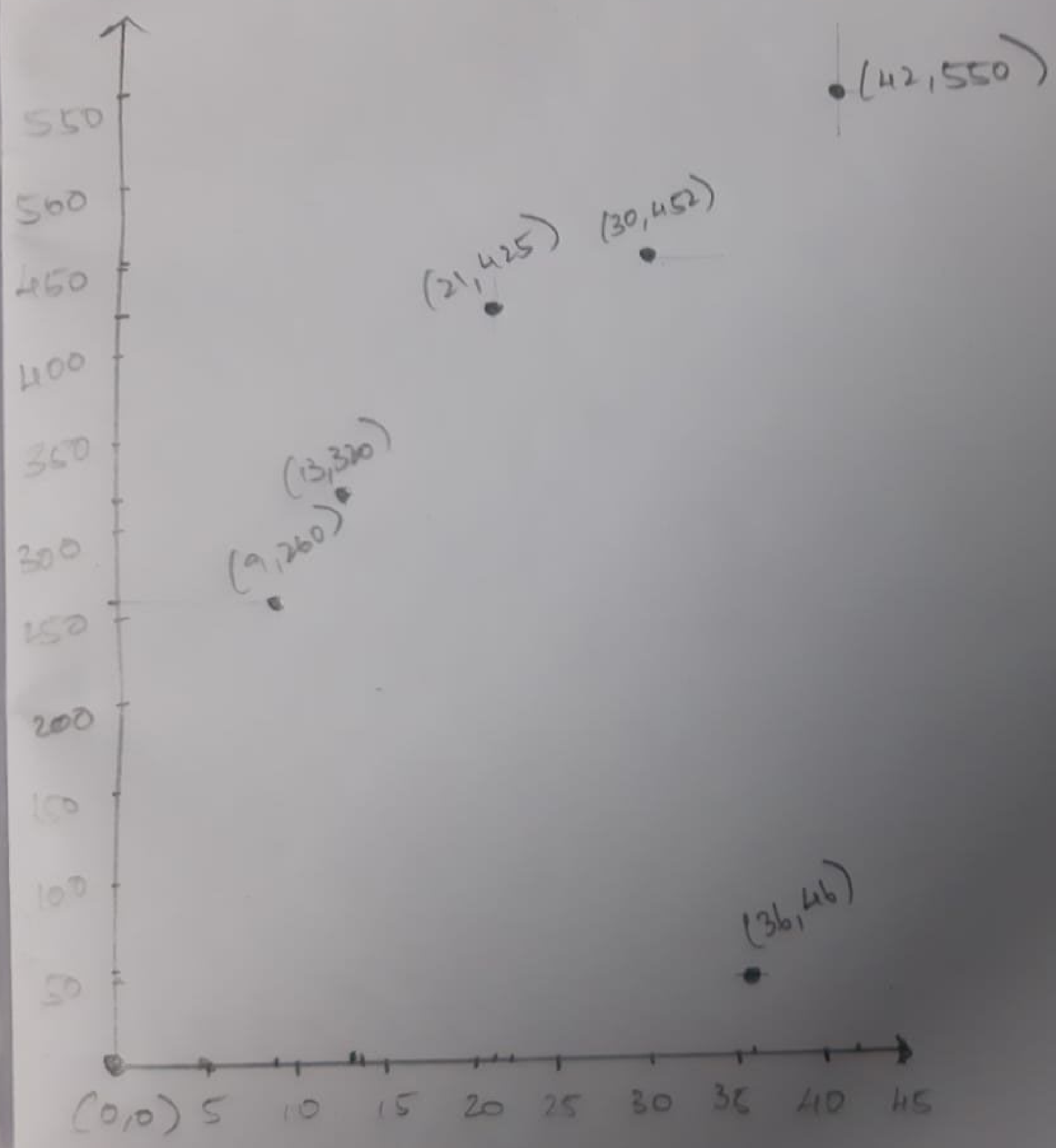
ii) From the graph,

clearly it is Positive Correlation

iii) To find best fit line,

$$\bar{x} = \frac{0+9+13+21+30+36+42}{7} \simeq 21.57$$

$$\bar{Y} = \frac{0+260+320+425+452+46+550}{7} \simeq 293.29$$


550
500
450
400
350
300
250
200
150
100
50
(0,0)
5
10
15
20
25
30
35
40
45
(9,260)
(13,320)
(21,425)
(30,452)
(42,550)
(36,46)

3) Interpretations,

a) Average Student Sleeps MOST on Friday

b) Average Student Sleeps LEAST on Thursday

c) The Sleep time varies from student to student MOST on Saturday

d) The Sleep times of students are very Similar (Low Variance) on Wednessday

e) HIGHEST Sleep time of a Student is on Sunday

f) LEAST Sleep time of a student is on Thursday.

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

$= \dfrac{(-21.57)(-293.29) + (-12.57)(-33.29) + (-8.57)(26.7\ldots) + (-0.57)(131.71) + (8.43)(158.71) + (14.43)(-247.2\ldots) + (20.43)(256.71)}{(-21.57)^2 + (-12.57)^2 + (-8.57)^2 + (-0.57)^2 + (8.43)^2 + (14.43)^2 + (20.43)^2}$

$= \dfrac{6326.2653 + 418.4553 - 228.9047 - 75.0747 + 1337.9253 - 3568.3947 + 5244.5853}{465.2649 + 158.0049 + 73.4449 + 0.3249 + 71.0649 + 208.2249 + 417.3849}$

$= \dfrac{9454.8571}{1394.7143} = 6.78$

y intercept $c = \bar{Y} - m\bar{X}$

$= 293.29 - 6.78 \times 21.57 \simeq 147$

$\therefore$ Best fit line $\Rightarrow$ $\underline{y = 6.78x + 147.}$

iv) $y = 400$ calories

As line $\Rightarrow$ $y = 6.78x + 147$,

Exercise time $= x = \dfrac{y - 147}{6.78} = \dfrac{400 - 147}{6.78}$

$= \underline{37.32}$

(3)

a)

b)

c)

d)

e)

f)

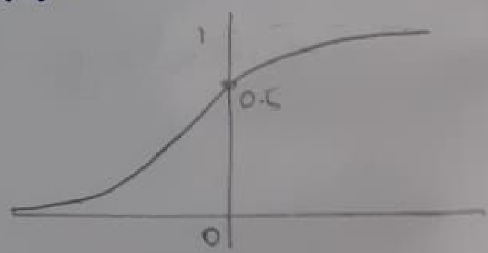
5. A)

Activation functions are needed in neural networks as they add some kind of non-linearity property to the network. As neural networks generally deal with complex relationships with data, these relationships CANNOT be accurately modelled using just linear relations.
So, activation functions are necessary to model non-linear complex relationships accurately.

Act fns,

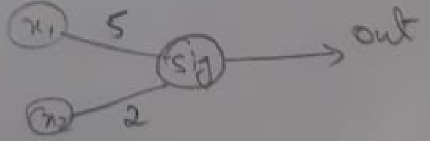
i) Sigmoid Activation Function

$$\text{Sigmoid}(x) = \frac{e^{x}}{1+e^{x}}$$



Trace,

Consider a neuron,



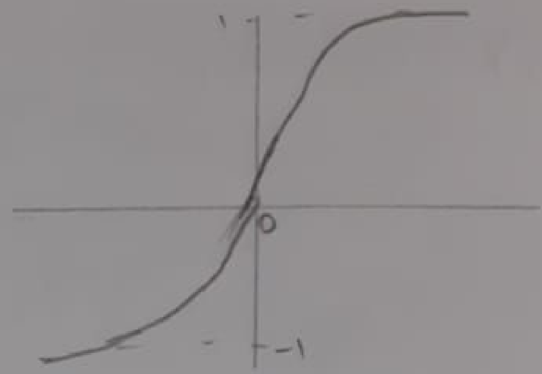
using sigmoid act fn,

for input $x_1 = 1$, $x_2 = 2$,

$$\text{out} = \frac{e^{1\times5+2\times2}}{1+e^{1\times5+2\times2}} = \frac{e^{9}}{1+e^{9}} \simeq 0.99988$$

ii) Tanh

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



Trace for same scenario as (i),

$x_1 = 1$, $x_2 = 2$, $w_1 = 5$, $w_2 = 2$,

$$out = \frac{e^{9} - e^{-9}}{e^{9} + e^{-9}} = 0.9999$$

iii) Rectified Linear Unit

$$\text{ReLu}(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Trace for same scenario as (i),

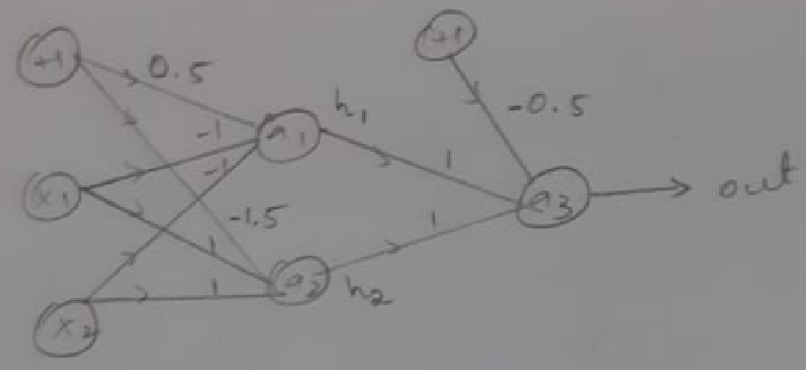
$x_1 = -1$, $x_2 = 2$, $w_1 = 5$, $w_2 = 2$,

$out \Rightarrow x_1 w_1 + x_2 w_2 = -5 + 4 = -1$

as $-1 < 0$,

$out = 0$,,

B)



Act fn, $f(x) = \begin{cases} 1 & , x >= 0 \\ 0 & , x < 0 \end{cases}$

<u>Input Layer,</u>

inputs = x_1, x_2 bias $b_1 = +1$

<u>Hidden Layer,</u>

At neuron a_1,

output of $a_1 = f(w_{11}x_1 + w_{12}x_2 + w_{1b}b_1)$

as $w_{11} = -1$, $w_{12} = -1$, $w_{1b} = 0.5$, $b_1 = +1$,

$= f(-x_1 - x_2 + 0.5)$

$\therefore$ h_1 = output of a_1 = $f(0.5 - (x_1 + x_2))$

At neuron a_2,

output of $a_2 = f(w_{21}x_1 + w_{22}x_2 + w_{2b}b_1)$

as $w_{21} = 1$, $w_{22} = 1$, $w_{2b} = -1.5$, $b_1 = +1$,

$= f(x_1 + x_2 - 1.5)$

h_2 = output of a_2 = $f(x_1 + x_2 - 1.5)$

Output Layer.

$h_1 = f(0.5 - (x_1 + x_2))$

$h_2 = f(x_1 + x_2 - 1.5)$

$b_2 = +1, \quad \omega_{31} = +1, \quad \omega_{32} = +1, \quad \omega_{3b} = -0.5$

$\therefore$ At output neuron a_3,

output $= f(\omega_{31} h_1 + \omega_{32} h_2 + \omega_{3b} b_2)$

$= f(h_1 + h_2 - 0.5)$

output $= f\big(f(0.5 - x_1 - x_2) + f(x_1 + x_2 - 1.5) - 0.5\big)$

Now, if we consider,

$x_1 + x_2 = X$ (say)

① Then, if $X \leq 0.5$,

then, ~~output~~ $f(0.5 - X) = 1$ as $0.5 - X \geq 0$.
as $X \leq 0.5$.

and $f(X - 1.5) = 0$ as $X - 1.5 < 0$.

$\therefore$ output $= f(1 + 0 - 0.5) = f(0.5) = \boxed{1}$

② If $0.5 < X < 1.5$

then $f(0.5 - X) = 0$ as $X > 0.5$
So, $0.5 - X < 0$

and $f(X - 1.5) = 0$ as $X < 1.5$
so, $X - 1.5 < 0$

output $= f(0 + 0 - 0.5) = f(-0.5) = \boxed{0}$
as $-0.5 < 0$.

③ If $x \geq 1.5$,

$f(0.5 - x) = 0$ as $x > 0.5$

So, $0.5 - x < 0$

$f(x - 1.5) = 1$ as $x \geq 1.5$,

So, $x - 1.5 \geq 0$

output $= f(0 + 1 - 0.5) = f(0.5) = \boxed{1}$

So, we can infer that,

Neural Network is simulating a function,

$$g(x_1, x_2) = \begin{cases} 1, & x_1 + x_2 \leq 0.5 \\ 0, & 0.5 < x_1 + x_2 < 1.5 \\ 1, & x_1 + x_2 \geq 1.5 \end{cases}$$

$g(x_1, x_2)$ is a function that returns 0 if x_1, x_2 sum lies exclusively between 0.5 and 1.5 and it returns 1 otherwis

3\. A) $\overset{P_1}{(2,10)}$, $\overset{P_2}{(2,5)}$, $\overset{P_3}{(8,4)}$, $\overset{P_4}{(5,8)}$, $\overset{P_5}{(7,5)}$, $\overset{P_6}{(6,4)}$,
$\underset{P_7}{(1,2)}$, $\underset{P_8}{(4,9)}$, $\underset{P_9}{(8,6)}$, $\underset{P_{10}}{(6,7)}$

i) Single Link Strategy,

First we compute distance matrix and we find least dist b/w 2 points.

We name the points as $P_1, P_2, P_3 \ldots P_{10}$. respectively.

Upon computing,

min dist = 1.414

b/w points, (P_3, P_5), (P_4, P_8), (P_4, P_{10}), (P_5, P_6), (P_5, P_9)

So, we combine into clusters, c_1 and c_2.

$\overset{c_1}{(P_3\ P_5\ P_6\ P_9)}$ and $\overset{c_2}{(P_4\ P_8\ P_{10})}$

Then we recalculate Proximity matrix such that,

$$dist(c_1, P_i) = \min\big(dist(P_3, P_i), dist(P_5, P_i), dist(P_6, P_i), dist(P_9, P_i)\big)$$

similarly for c_2.

Also, Branch length is calculated for each cluster.

Branch Length of c_1 = 1.414 / 2 = 0.707

Branch Length of c_2 = 1.414 / 2 = 0.707

6) Heirarchical Clustering vs K Means

i) Time Complexity:

K Means - $O(n)$ Linear

Heirarchical - $O(n^2)$ Quadratic

So, KMeans can handle big data well and executes faster than heirarchical.

ii) Reproducability:

In KMeans, we start with random choice of clusters and so results may vary with each different run. So, results are non-reproducible in KMeans.

But, in heirarchical clustering, as there is no random aspect, results are reproducible.

iii) Number of Clusters:

KMeans requires prior knowledge of 'K' number of clusters.

But, heirarchical doesnt require such knowledge and so we can decide and get clusters for any no. of clusters by interpreting the dendrogram.