Principles of Machine Learning

Unit 3

Probabilistic Learning, Learning with Trees

Questions on Baye's Theorem

CAT-2 Consider the following dataset and apply a probabilistic model using Baye's Theorem to classify a species as M or H

S.no	color	legs	height	smelly	species
1	white	3	short	Y	M
2	green	2	tall	N	M
3	green	3	short	Y	M
4	white	3	short	Y	M
5	green	2	short	N	H
6	white	2	tall	N	H
7	white	2	tall	N	H
8	white	2	short	Y	H

Identify the species when

$X = \{$ color = green, legs = 2, height = tall, smelly = no $\}$

Ans

Bayes Rule: $P(A|B) = \dfrac{P(B|A)\,P(A)}{P(B)}$

$P(M) = 1/2$

$P(H) = 1/2$

color

color	M	H
green	1/2	1/4
white	1/2	3/4
	=1	=1

$$P(\text{green}\mid M) = \frac{P(\text{green} \cap M)}{P(M)}$$

legs

legs	M	H
2	1/4	1
3	3/4	0
	=1	=1

height

height	M	H
short	3/4	2/4
tall	1/4	2/4

smelly

smelly	M	H
yes	3/4	1/4
no	1/4	3/4

$$P(M\mid X)$$

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

$$= P(X\mid M) \times P(M)$$

$$= P(\text{Green}\mid M) \times P(2\text{ legs}\mid M) \times P(\text{tall}\mid M)$$

$$\times P(\text{notsmelly}\mid M) \times P(M)$$

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

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

$$= 3.9\times10^{-3} = \underline{0.0039}$$

$P(H|X)$

$= P(X|H)P(H)$

$= P(\text{green}|H) \times P(\text{2legs}|H) \times P(\text{tall}|H) \times P(\text{not smelly}|H)$

$\times P(H)$

$= \frac{1}{4} \times 1 \times \frac{2}{4} \times \frac{3}{4} \times \frac{1}{2} = \underline{\underline{0.046}}$

$\therefore$ The species is $\underline{\underline{H}}$

Q2 Use a Naive Bayes classifier to find the suitable activity for the given test case.

Deadline = Near

Party = yes

Lazy = yes

	Deadline (D)	Party (P)	Lazy (L)	Activity A
1	urg	y	y	Party
2	urg	N	y	Study
3	near	y	y	Party
4	none	y	N	Party
5	none	N	y	Pub
6	none	y	N	Party
7	near	N	N	Study
8	near	N	y	TV
9	near	y	y	Party
10	urg	N	N	Study

Deadline

n(Party) = 5

n(Study) = 3

n(TV) = 1

n(Pub) = 1

	Party	Study	Pub	TV
None	2/5	0	1	0
Near	2/5	1/3	0	1
Urgent	1/5	2/3	0	0

Party

	Party	Study	Pub	TV
Y	1	0	1	0
N	0	1	0	1

Laziness

	Party	Study	Pub	TV
Y	3/5	1/3	1	1
N	2/5	2/3	0	0

Test case: Near, No (Party), Yes (Lazy)

$\boxed{\text{Party}} = \frac{2}{5} \times 0 \times \frac{3}{5} \times \frac{1}{2} = 0$

$\boxed{\text{Study}} = \frac{1}{3} \times 1 \times \frac{1}{3} \times 3/10 = \frac{1}{30}$

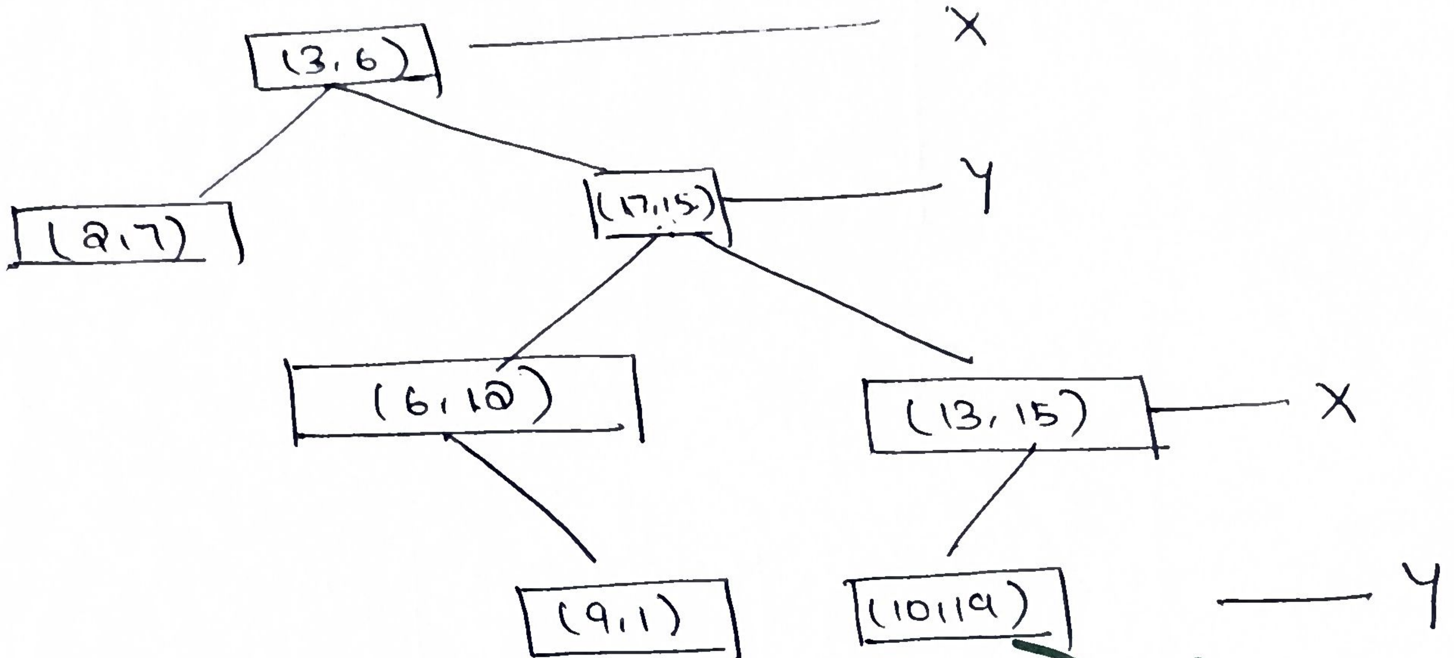
$\boxed{\text{Pub}} = 0$

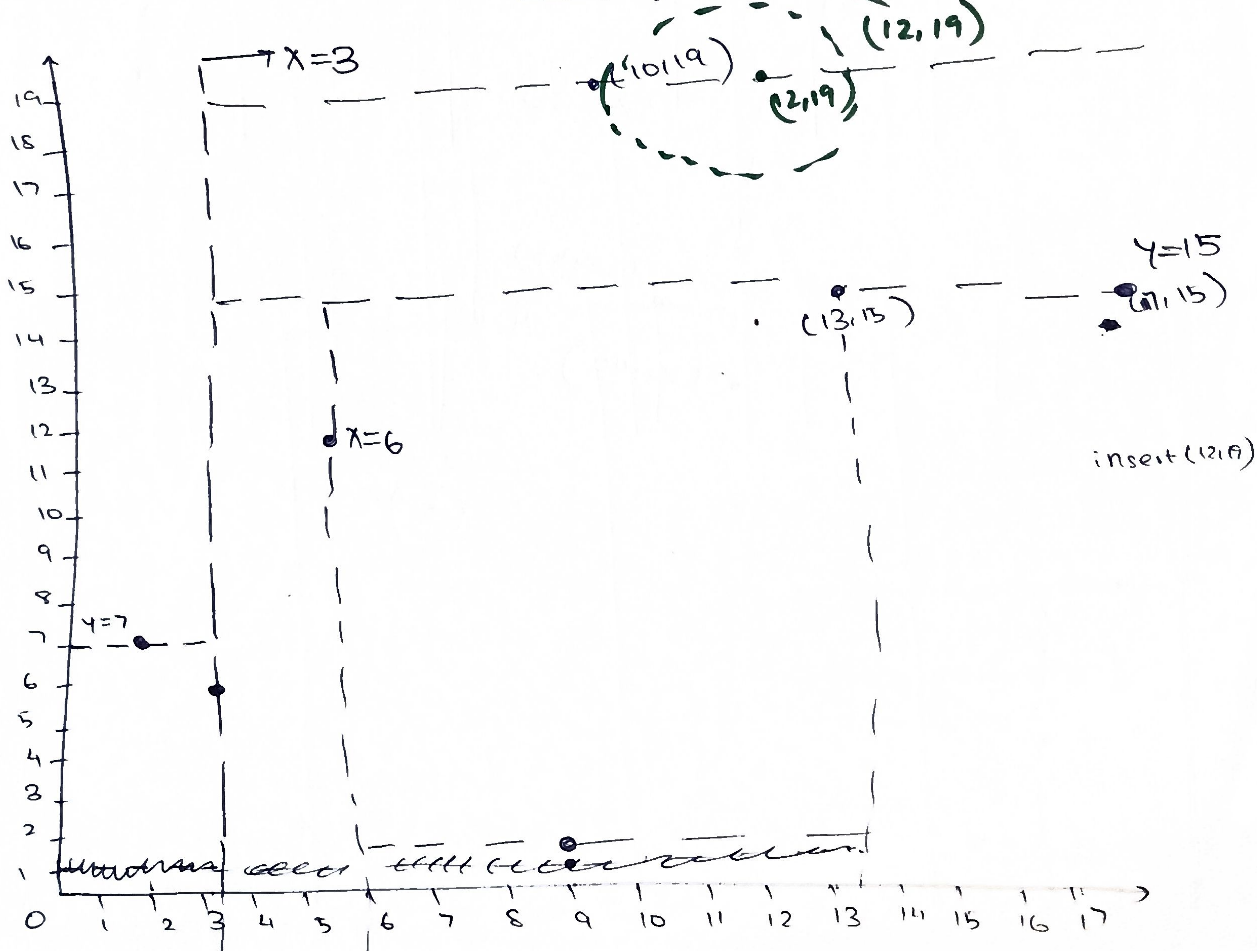
$\boxed{\text{TV}} = 1 \times 1 \times 1 \times \frac{1}{10} = \frac{1}{10}$

$\boxed{\text{action} = \text{TV}}$

KD Trees

1. PPT Q Create a KD Tree for the points (3,6), (17,15), (13,15), (6,12), (9,1), (2,7), (10,19) and search for (12,19) and (10,19)

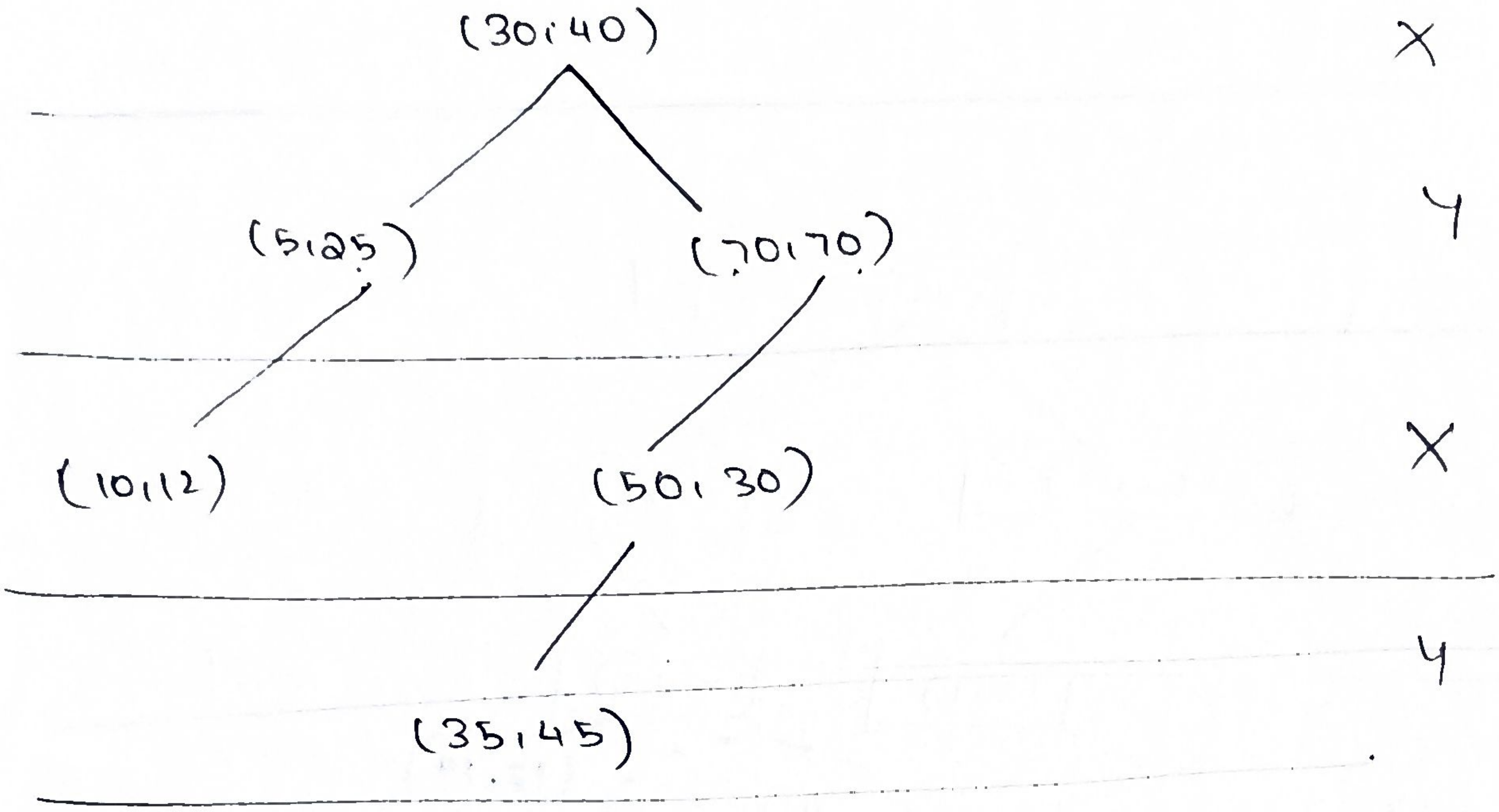


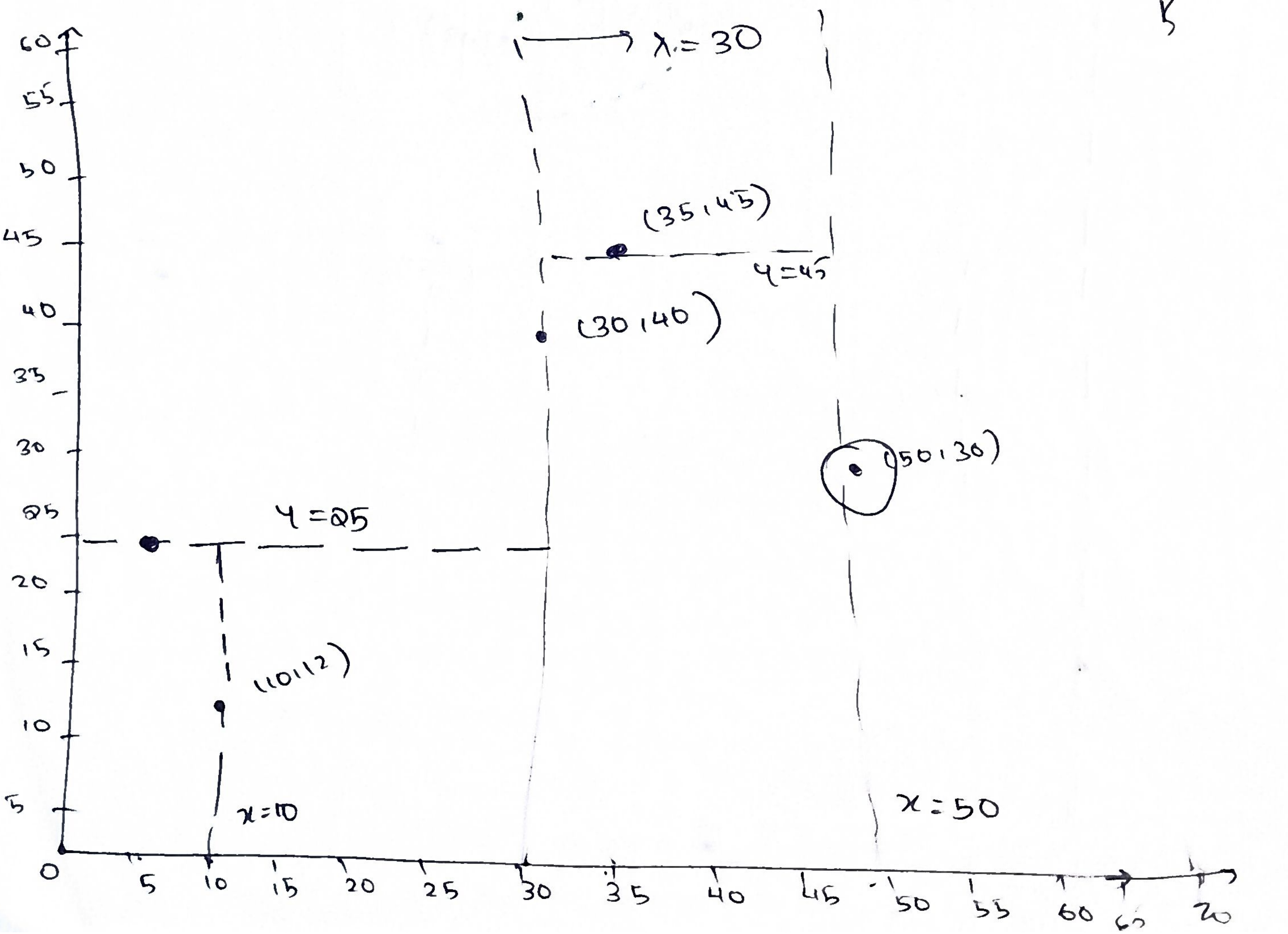


2. $\boxed{\text{CAT-2}}$ Construct a KD Tree with the following pts.

(30, 40) (5, 25) (10, 12) (70, 70) (50, 30) (35, 45)

Search for (50, 30)





KNN Numerical

1. For the given dataset, use KNN to find the class of height = 170, weight = 57.

Height	Weight	Class
167	51	U
182	62	N
176	69	N
173	64	N
172	65	N
174	56	U
169	58	N
173	57	N
170	55	N

<u>Ans</u> Compute the Euclidean distance between the given point & every other point.

Height	Weight	class	distance	
167	51	U	6.7	5
182	62	N	13	8
176	69	N	13.4	9
173	64	N	7.6	6
172	65	N	8.2	7
174	56	U	4.1	4
169	58	N	1.4	1
173	57	N	3	3
170	55	N	2	2

<u>Rank them</u>

when $K = 1 \Rightarrow N$

$K = 2 \Rightarrow \{N, N\} \Rightarrow N$

$K = 3 \Rightarrow \{N, N, N\} \Rightarrow N$

$K = 4 \Rightarrow \{U, N, N, N\} \Rightarrow N$

(from majority voting)

Decision Trees

CAT2 Consider the dataset given below and construct the decision tree model using information gain & entropy as the impurity metric.

Height	Hair	Eyes	Attractive?
S	B	Brown	No
T	D	Brown	No
T	B	Blue	Yes
T	D	Blue	No
S	D	Blue	No
T	R	Blue	Yes
T	B	Brown	No
S	B	Blue	Yes

$P(yes) = 3/8$

$P(no) = 5/8$

$$E(S) = -p_{yes} \log p_{yes} - p_{no} \log p_{no}$$

$$= -\frac{3}{8} \log \frac{3}{8} - \frac{5}{8} \log \frac{5}{8}$$

$\cancel{0.5 + 0.42}$ $= \underline{0.954}$

~~Height~~ $= \cancel{0.954} - \{ \cancel{\frac{3}{8} \log \frac{3}{8}} - 1/3 \log 1/3 -$

Height

$$IG = 0.954 - \left\{ \frac{3}{8} \times \left(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \right) \right.$$

$$\left. - \frac{5}{8} \times \left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} \right) \right\}$$

$$= \underline{0.069}$$

Hair

$$IG = 0.954 - \left\{ \frac{4}{8} \left(-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} \right) \right.$$

$$- \frac{1}{8} \left(-\frac{1}{1}\log\frac{1}{1} - \frac{0}{1}\log\frac{0}{1} \right)$$

$$\left. - \frac{3}{8} \left(-\frac{3}{3}\log\frac{3}{3} - \frac{1}{1}\log\frac{1}{1} \right) \right\}$$

$$= \underline{0.4544}$$

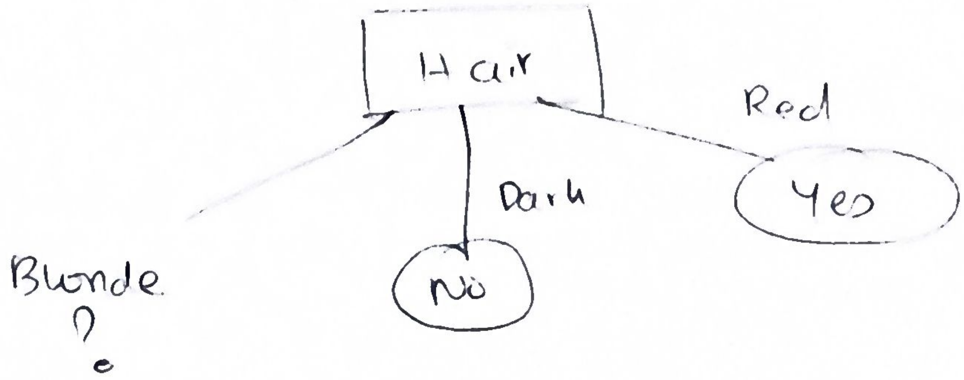
Eyes

$$0.954 \left\{ \frac{3}{8} \left(-\frac{0}{3}\log\frac{3}{3} - \frac{3}{3}\log\frac{3}{3} \right) \right.$$

$$\left. - \frac{5}{8} \left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} \right) \right\}$$

Hair = Root node

$$= \underline{0.3181}$$



Height	Hair	Eyes	Attractive
S	B	Brown	No
T	B	Blue	Yes
T	B	Brown	No
S	B	Blue	Yes

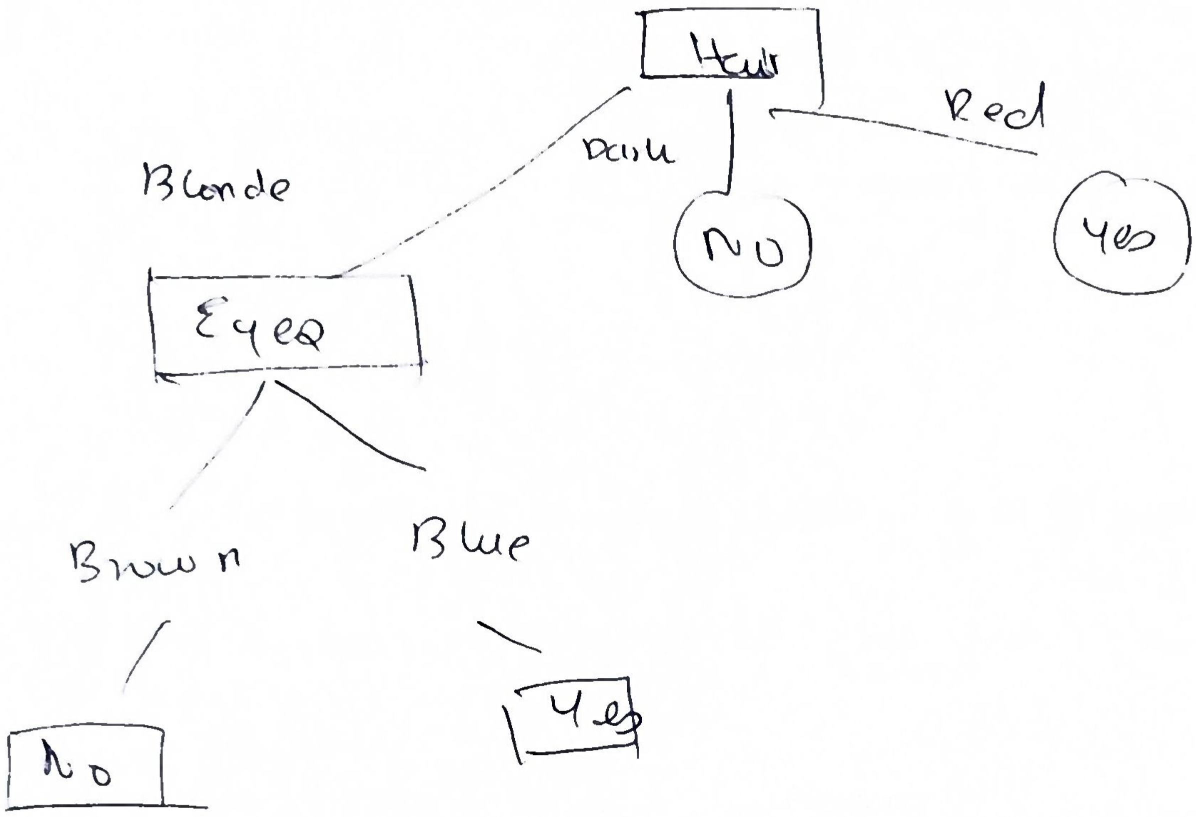
$$E(\text{Blonde}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}$$

IG(Blonde, Height)

$$= (E\ \text{blonde}) - \left\{ \frac{2}{4}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) - \frac{2}{4}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) \right\}$$

IG(Blonde, Eyes)

$$= E(\text{blonde}) - \left\{ \frac{2}{4}\left(-\frac{2}{2}\log\frac{2}{2} - 0\right) - \frac{2}{4}\left(-\frac{2}{2}\log\frac{2}{2} - 0\right) \right\}$$



Q2 Make use of the entropy metric and compute the impurity value for:

[5+, 5-], [0+, 10-], [8+, 2-]

(i) $[5+, 5-] = -\frac{5}{10}\log\frac{5}{10} - \frac{5}{10}\log\frac{5}{10} = 1$ ⇒ highly impure

(ii) $[0+, 10-] = \frac{0}{10}\log 0 - \frac{10}{10}\log 10 = 0$ ⇒ pure

(iii) $[8+, 2-] = \frac{8}{10}\log\frac{8}{10} - \frac{2}{10}\log\frac{2}{10}$ ⇒ impure

Q3 The probability of 5 events are P(first) = 0.5

P(second) = P(third) = P(fourth) = P(fifth) = 0.125

Calculate entropy.

Ans: $-0.5\log 0.5$
$-\ 4(0.125\log 0.125)$