Data Mining (CS4038)

Quiz No.3

Roll No: ____________________ **Section:** __________ **Date:11-03-2024**

Question No.1 **(5+5=10 marks)**

Consider a training dataset that contains 100 positive examples and 300 negative examples. For each of the following rules:

R1: A → + (covers 50 positive and 70 negative examples)

R2: A → + (covers 80 positive and 40 negative examples)

Determine which is the best rule according to i) accuracy and ii) Foil's Information Gain.

Note: Perform each step to secure marks.

Accuracy $R_1 = \frac{50}{50+70} \times 100\%$

$= 41.6\%$

$= 0.41$

Accuracy $R_2 = \frac{80}{80+40} \times 100\%$

$= 66.67\%$ R_2 is better.

Foil info gain

$= P_1 = P_1 \left[\log_2\left(\frac{P_1}{P_1+n_1}\right)\right] - \log_2\left(\frac{P_0}{P_0+n_0}\right)$

$= 36.85$

$R_2 = 80 \times \left(\log_2\left(\frac{80}{120}\right) - \log_2\left(\frac{100}{400}\right)\right)$

$= 113.28$

R_2 is better

Question No. 2 By Median along x, y **(8+2 =10 marks)**

Construct the Kdtree for the given dataset step by step showing the splitting criteria at each step:

Point	X	Y
P1	3	6
P2	17	15
P3	13	15
P4	6	12
P5	9	5
P6	5	7
P7	3	4

Also perform a nearest neighbor search using the constructed KD-tree and determine the closest point to the query point Q (8,8).

Along x

3	6
3	4
5	7
Median → 6	12
9	5
13	15
17	15

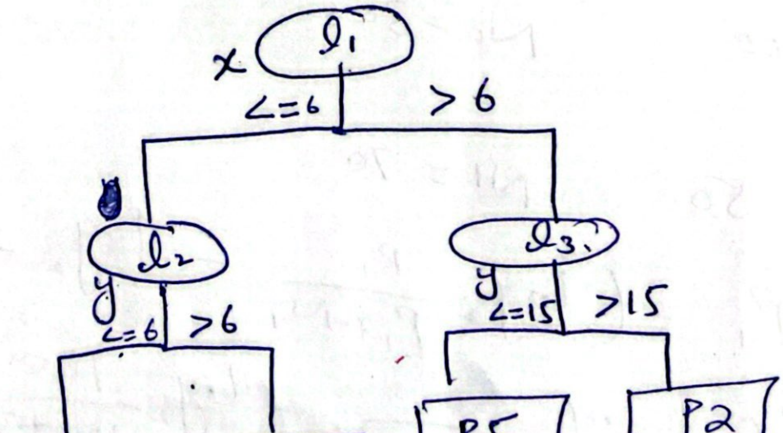
⇒

✓ <= Along y >

3	4
3	6 Median
5	7

9	5
13	15 ← Median
17	15

+8



Q (8,8)

Its nearest neighbors is P5 with path $l_1 - l_3 - P5$.

+2