Q1

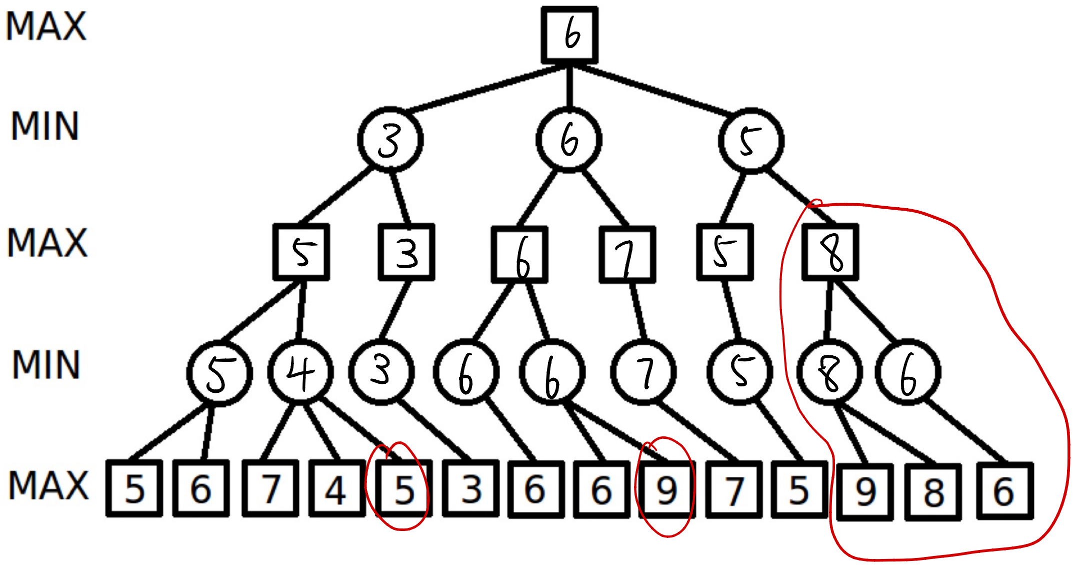
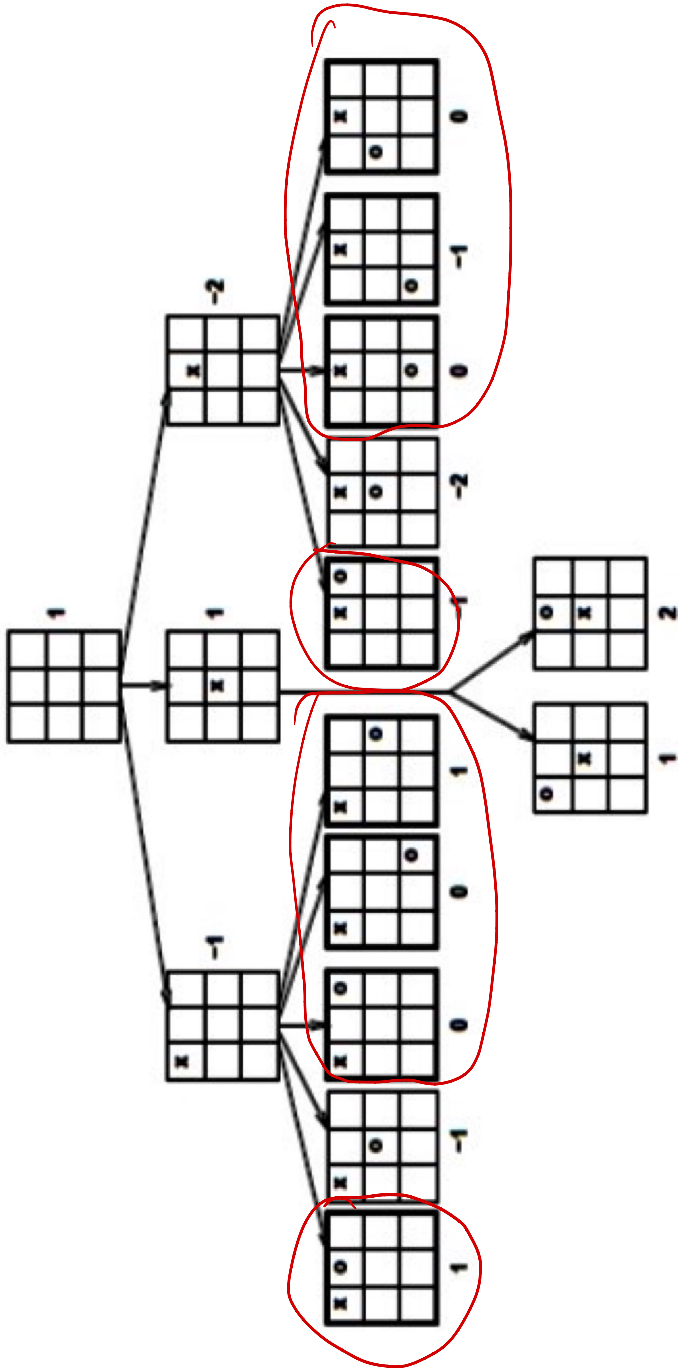


Figure 1: Game tree

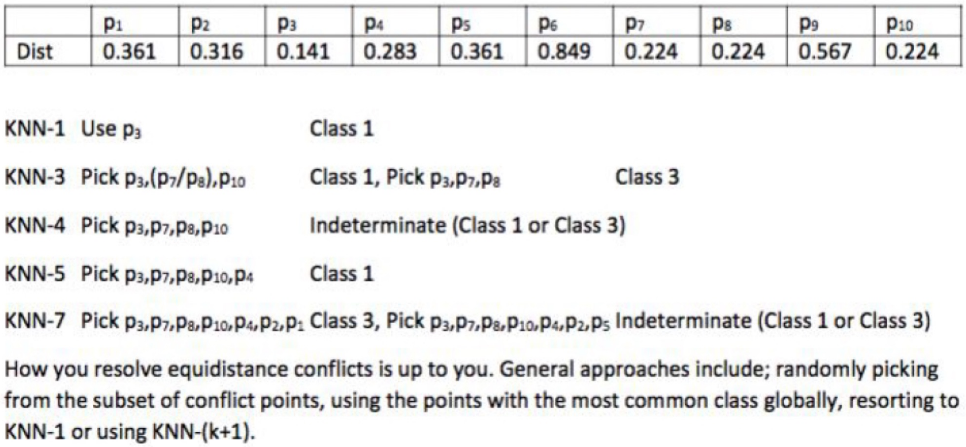
No need to search the nodes in red circles

Q2



Prune the circled nodes

Q3



	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
Dist	0.361	0.316	0.141	0.283	0.361	0.849	0.224	0.224	0.567	0.224

KNN-1 Use p_3 Class 1

KNN-3 Pick p_3,(p_7/p_8),p_{10} Class 1, Pick p_3,p_7,p_8 Class 3

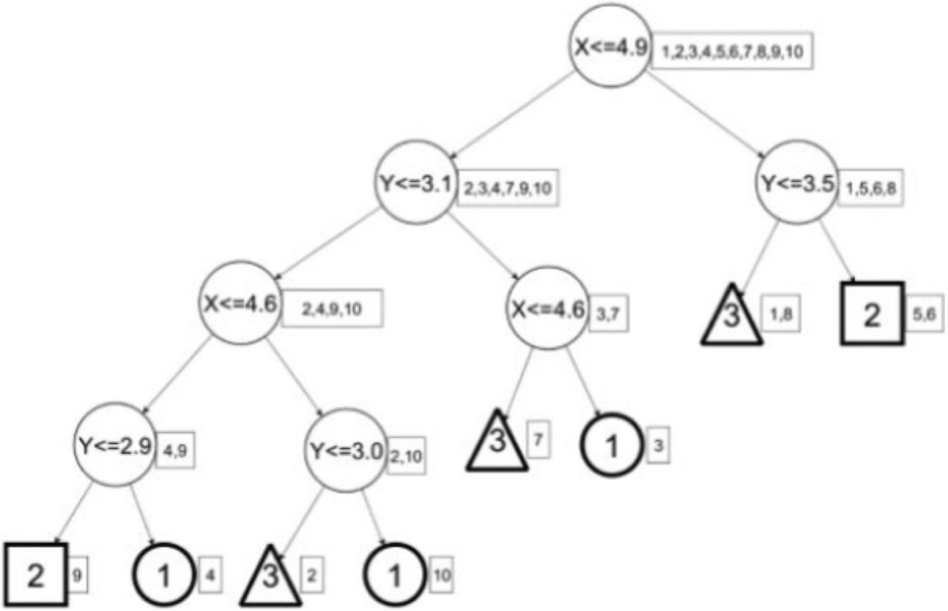
KNN-4 Pick p_3,p_7,p_8,p_{10} Indeterminate (Class 1 or Class 3)

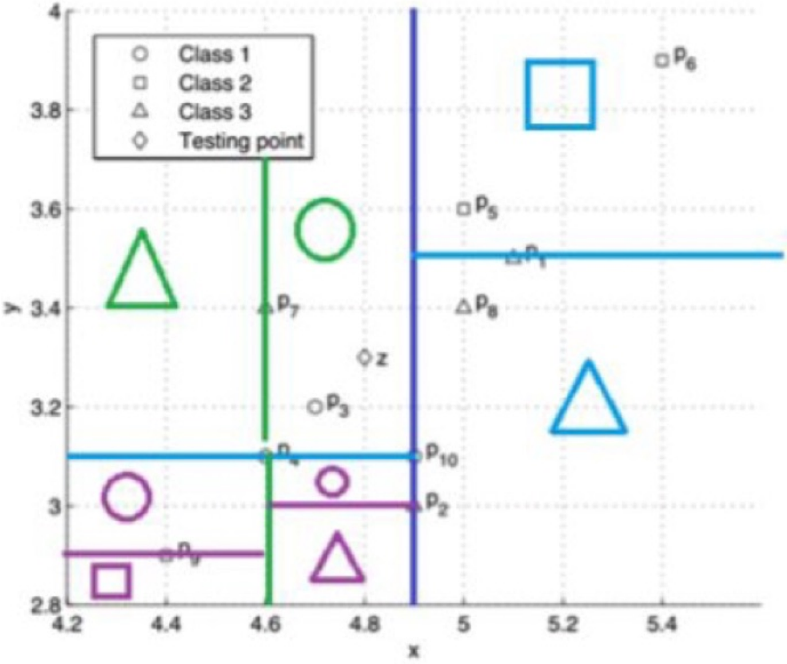
KNN-5 Pick p_3,p_7,p_8,p_{10},p_4 Class 1

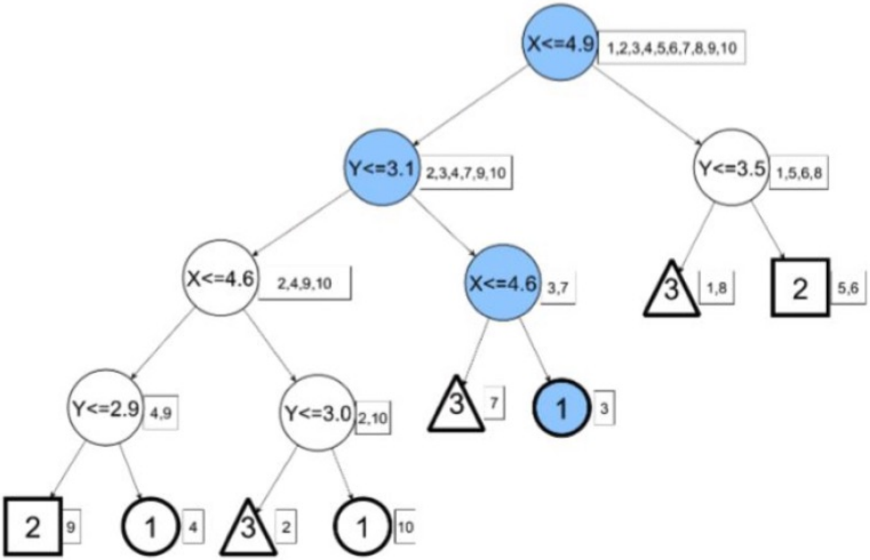
KNN-7 Pick p_3,p_7,p_8,p_{10},p_4,p_2,p_1 Class 3, Pick p_3,p_7,p_8,p_{10},p_4,p_2,p_5 Indeterminate (Class 1 or Class 3)

How you resolve equidistance conflicts is up to you. General approaches include; randomly picking from the subset of conflict points, using the points with the most common class globally, resorting to KNN-1 or using KNN-(k+1).

Q4




Class 1
Class 2
Class 3
Testing point
4
3.8
3.6
3.4
3.2
3
2.8
y
4.2
4.4
4.6
4.8
5
5.2
5.4
x
p6
p5
p1
p7
p8
z
p3
p4
p10
p2
p9


X<=4.9
1,2,3,4,5,6,7,8,9,10
Y<=3.1
2,3,4,7,9,10
Y<=3.5
1,5,6,8
X<=4.6
2,4,9,10
X<=4.6
3,7
3
1,8
2
5,6
Y<=2.9
4,9
Y<=3.0
2,10
3
7
1
3
2
9
1
4
3
2
1
10

$$\mathrm{H}(\mathcal{S}) = -\sum p_i \log_2 p_i \,,$$

$$IG(T,a) = H(T) - \sum_{v \in vals(a)} \frac{|\{\mathbf{x} \in T | x_a = v\}|}{|T|} \cdot H(\{\mathbf{x} \in T | x_a = v\})$$

Depth 0:

$$H(Root) = -\left(\frac{11}{16}\log_2\frac{11}{16} + \frac{5}{16}\log_2\frac{5}{16}\right) \approx 0.896$$

Split time:

[[1,9],[2,4,5,6,7,8,10,13,14,15,16],[3,11,12]]

$$H(\boldsymbol{x} = M) = -(1\log_2 1 + 0) = 0$$

$$H(\boldsymbol{x} = A) = -\left(\frac{7}{11}\log_2\frac{7}{11} + \frac{4}{11}\log_2\frac{4}{11}\right) \approx 0.946$$

$$H(\boldsymbol{x} = N) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \approx 0.918$$

$$IG(Time) = 0.896 - \left(\frac{2}{16}0 + \frac{11}{16}0.946 + \frac{3}{16}0.918\right) \approx 0.074$$

Split Match:

[2,6,7,8,10,15,16],[1,5,9,12,13,14],[3,4,11]

$$H(\boldsymbol{x} = G) = -\left(\frac{6}{7}\log_2\frac{6}{7} + \frac{1}{7}\log_2\frac{1}{7}\right) \approx 0.381$$

$$H(\boldsymbol{x} = M) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) \approx 1.0$$

$$H(\boldsymbol{x} = F) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \approx 0.918$$

$$IG(Match) = 0.896 - \left(\frac{7}{16}0.381 + \frac{6}{16}1.0 + \frac{3}{16}0.918\right) \approx 0.182$$

Split Surface:

[1,6,9,14],[3,7,8,11,15],[2,5,10,13,16],[4,12]

$$H(\boldsymbol{x} = G) = -(1\log_2 1 + 0) = 0$$

$$H(\boldsymbol{x} = H) = -(1\log_2 1 + 0) = 0$$

$$H(\boldsymbol{x} = C) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0.971$$

$$H(\boldsymbol{x} = M) = -(0 + 1\log_2 1) = 0.0$$

$$IG(Surface) = 0.896 - \left(\frac{5}{16}0.971\right) \approx 0.593$$

Choose Surface as it has the highest information gain

Depth 2:

All non-clay surface games are leaf nodes (entropy of zero). Need to split clay further (entropy is non-zero). The set is now [2,5,10,13,16].

$$H(Surface = Clay) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0.971$$

Split time:

[],[2,5,10,13,16],[]

Omitted as this split trivially gives us nothing

Split Match:

[2,10,16],[5,13],[]

$$H(\boldsymbol{x} = G) = -\left(\tfrac{2}{3}\log_2\tfrac{2}{3} + \tfrac{1}{3}\log_2\tfrac{1}{3}\right) \approx 0.918$$

$$H(\boldsymbol{x} = M) = -(0 + 1\log_2 1) = 0.0$$

$$H(\boldsymbol{x} = F) = 0.0$$

$$IG(Match) = 0.971 - \left(\tfrac{3}{5}0.918 + 0\right) \approx 0.420$$

Split Surface:

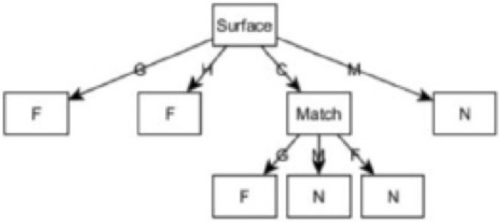
Discrete values, skip

Choose Match as it has the highest information gain

Depth 3:

Set is now [2,10,16], we cannot split this further so we assign the most probable value at this node.

Assign most probable outcome of parent to the node F as it has no elements.



I also mentioned two other useful equations:

$\log_a k = \frac{\log_b k}{\log_b a}$, eg. $\log_2 42 = \frac{\ln 42}{\ln 2} = 5.39$

And:

$H_b(S) = -\sum p_i \log_b p_i$, where b is the number of possible outcomes (2 for binary events).

Q6

Red: 40
Green: 40
Total: 80

H(Total) = 1
Split length by <= 5.5 -> setosa
38 on L>5.5 side, 42 on L<=5.5

a//

H(L > 5.5) = -((37/38) * log(37/38) + (1/38) * log(1/38))

H(L <= 5.5) = -((3/42) * log(3/42) + (39/42) * log(39/42))

IG = 1 - 38/80 * H(L > 5.5) + 42/80 * H(L <= 5.5) = 1 - 0.02510391101 - 0.05866983448 ~= 0.918

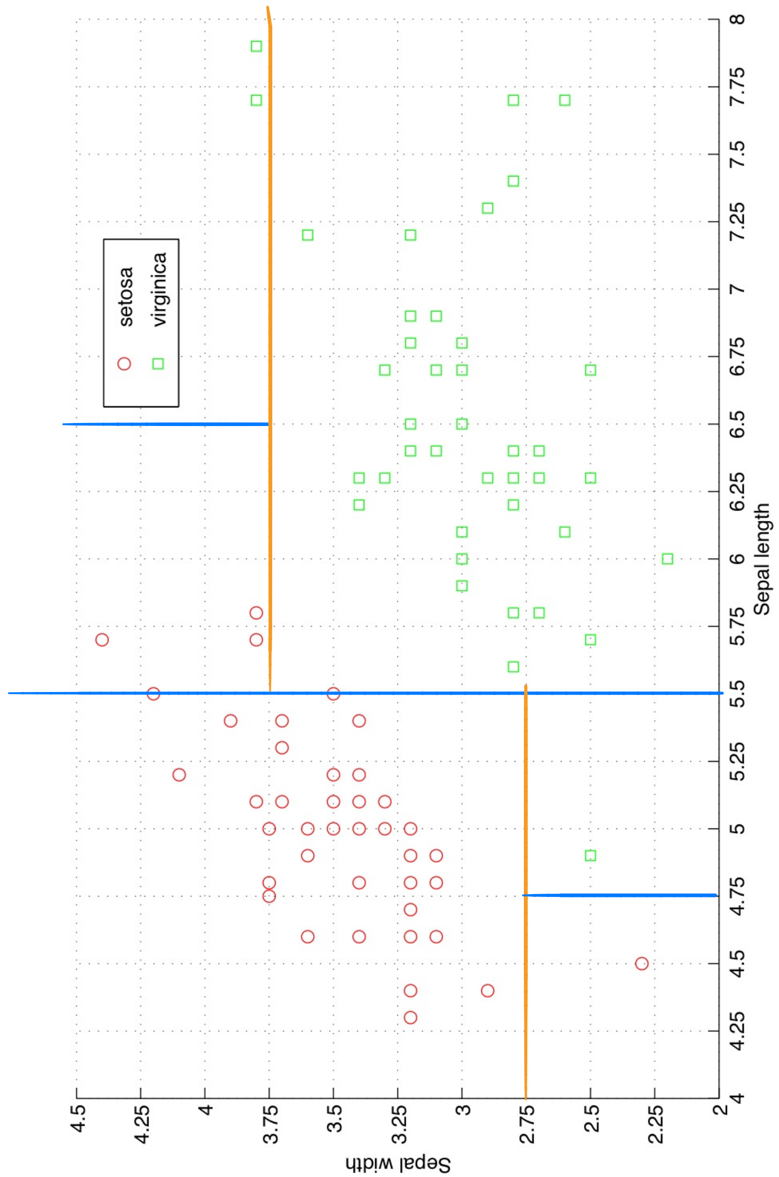
Splitting on 5.75 will give the same IG

b//

Same procedure, on L <= 5.5 it would be 3 and on L > 5.5 it would be 3.75

c

5 splits


setosa
virginica
Sepal length
Sepal width
4
4.25
4.5
4.75
5
5.25
5.5
5.75
6
6.25
6.5
6.75
7
7.25
7.5
7.75
8
2
2.25
2.5
2.75
3
3.25
3.5
3.75
4
4.25
4.5