Assignment 3

1. consider attribute a_1, The corresponding counts and probabilities are

a_1	+	-
T	3	1
F	1	4

The entropy for a_1 is

$$\frac{4}{9}\left[-(3/4)\log_2(3/4) - (1/4)\log_2(1/4)\right] + \frac{5}{9}\left[-(1/5)\log_2(1/5) - (4/5)\log_2(4/5)\right] = 0.7616$$

The ~~info~~ entropy of the training example is

$$-P(+)\log_2 P(+) - P(-)\log_2 P(-) = -4/9\log_2(4/9) - 5/9\log_2(5/9)$$

$$= 0.9911$$

∴ The information gain for a_1 is $0.9911 - 0.7616$

$= 0.2294$

For attribute a_2, the information gain is computed for every possible split

∴ for split 1 ~~ie~~ i.e $a_2(1) = 1$ & $a_3(2) = 3$

∴ The split point will $\frac{1+3}{2} = 2$

Entropy can be calculated as

~~$\frac{2}{9}\left[-(1/2)\log_2(1/2) - (1/2)\log_2(1/2)\right]$~~

$$\frac{1}{9}\left[-1\log_2(1) - 0\right] + \frac{8}{9}\left[\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right)\right]$$

$= 0.8482$

The entropy of the entire attribute a_2 will be

$$\frac{-4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right) = 0.9911$$

∴ The information gain will be 0.9911 - 0.8482 = 0.1429

For split 2 i.e $a_{3(2)} = 3$ & $a_{3(3)} = 4$

split point = $\frac{7}{2} = 3.5$

Entropy can be calculated as

$$\frac{2}{9}\left[-1\log_2(1/2) - 1\log_2(1/2)\right] + \frac{7}{9}\left[-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$$

= 0.9885

∴ The information gain will be 0.0026

For split 3 i.e $a_{3(3)} = 4$ & $a_{3(4)} = 5$

split point = $\frac{9}{2} = 4.5$

Entropy will be

$$\frac{3}{9}\left[-\frac{2}{3}\log_2(2/3) - \frac{1}{3}\log_2(1/3)\right] + \frac{6}{9}\left[-\frac{2}{6}\log_2(2/6) - \frac{4}{6}\log_2(4/6)\right]$$

= 0.9183

∴ The information gain will be 0.0728

For split 4 i.e $a_{3(4)} = 5$ $a_{3(5)} = 5$ & $a_{3(6)} = 6$

split point = $\frac{11}{2} = 5.5$

Entropy will be

$$\frac{5}{9}\left[-\frac{2}{5}\log\frac{2}{2s} - \frac{3}{5}\log_2\frac{3}{5}\right] + \frac{4}{9}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$$

= 0.9830

∴ The information gain will be 0.0072

For split point 6 i.e $a_{3(6)} = 6$ & $a_{3(7)} = a_{3(8)} = 7$

split point = $\frac{13}{2} = 6.5$

∴ The entropy will be

$$\frac{6}{9}\left[\frac{-3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right] + \frac{3}{9}\left[\frac{-1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right]$$

$= 0.9728$

∴ The information gain will be 0.0183

For split 6 $a_2(7) = 9_3(8) = 7$ & $a_3(9) = 8$

Split point: $\frac{7+8}{2} = 7.5$

Entropy will be

$$\frac{8}{9}\left[\frac{-4}{8}\log_2\frac{4}{28} - \frac{4}{8}\log_2\frac{4}{28}\right] + \frac{1}{9}\left[0 - 1\log_2(1)\right]$$

$= 0.888$

∴ The information gain will be 0.1022

This can be summarized in the following table

a_2	Class label	Split point	Entropy	Info gain
1	+	2	0.8484	0.1427
3	-	3.5	0.9885	0.002
4	+	4.5	0.9183	0.0728
5	-	5.5	0.9839	0.0072
5	-	5.5	0.9728	0.0183
6	+	6.5	0.9728	0.0183
7	+	7.5	0.8888	0.1022
7	-	7.5	0.888	0.1022

The best split for attribute a_2 occurs at split point 2

By comparing the information gain of a_1 & a_2 we can conclude that a_1 produces the best split

2) Instance should not be used as another attribute since attribute instance has no predictive power

Problem 2

1. The contingency table after splitting on attribute A will be

A	+	-
T	20	30
F	15	35

~~The overall entropy before~~

The overall gini index before splitting would be

$$1 - \left(\frac{35}{100}\right)^2 - \left(\frac{65}{100}\right)^2 = 1 - 0.1225 - 0.4225 = 0.455$$

The gini index for attribute A will be

$$\frac{50}{100}\left[1 - \left(\frac{20}{50}\right)^2 - \left(\frac{30}{50}\right)^2\right] + \frac{50}{100}\left[1 - \left(\frac{15}{50}\right)^2 - \left(\frac{35}{50}\right)^2\right]$$

$= 0.45$

The contingency table after splitting on attribute B will be

B	+	-
T	15	20
F	20	45

The gini index for attribute B will be

$$\frac{35}{100}\left[1 - \left(\frac{15}{35}\right)^2 - \left(\frac{20}{35}\right)^2\right] + \frac{65}{100}\left[1 - \left(\frac{20}{65}\right)^2 - \left(\frac{45}{65}\right)^2\right]$$

$= 0.4947$

Since the gini index for a1 is smaller it produces better split.

For attribute A

A	+	-
T	20	30
F	15	35

cost	T	F
+	-1	100
-	0	-10

The cost of splitting is $20(-1) + 15(100) + 35(-10) = 1130$

For attribute B

B	+	-
T	15	20
F	20	45

cost	T	F
+	-1	100
-	0	0

The cost of splitting is $15(-1) + 20(0) + 20(100) + 45(0) = 1985$

since attribute A has the smallest cost we choose A for splitting

Problem 3

since there are 10 datapoints, the initial weight of each datapoint will be $D_1(1) = \cdots D_1(10) = \frac{1}{10}$

for 1st weak classifier H_1, points 9 and 10 are misclassified

$\therefore$ error $= 0.1 \times 2 = 0.2$

$$\alpha = \frac{1}{2} \log\left(\frac{1-0.2}{0.2}\right) = 0.301$$

The updated weights for the points not classified correctly would be

$D_2(i) = D_2(i) \exp(-\alpha_i y_i h_i(x_i)) = 0.1 e^{0.301 \times 1} = 0.135$

The updated weights for the correctly classified point is $D_2(i) = 0.1 \times e^{0.301 \times (-1)} = 0.074$

For 2nd weak classifier i.e H_2, points 1, 2, 3 and 8 are misclassified

error $= 0.1 \times 4 = 0.4$

$\alpha = \frac{1}{2}\log\left(\frac{1-0.4}{0.4}\right) = 0.088$

The updated weights for the points not classified correctly would be

$D_2(i) = 0.1e^{0.088 \times 1} = 0.109$

The updated weights for the correctly classified points would be

$D_2(i) = 0.1e^{0.088 \times (-1)} = 0.0915$

For 3rd weak classifier, i.e H_3, point 9 is misclassified

error $= 0.1 \times 1 = 0.1$

$\alpha = \frac{1}{2}\log\left(\frac{1-0.1}{0.1}\right) = 0.477$

The updated weights for the points not classified correctly would be

$D_3(i) = 0.1e^{0.477 \times (1)} = 0.161$

The updated weights for the correctly classified points would be

$D_3(i) = 0.1e^{0.477 \times (-1)} = 0.062$

2. For the 1st weak classifier the data instances that would be reweighted would be 9, 10. For the 2nd weak classifier the points 1, 2, 3 and 8 would be reweighted whereas for the 3rd weak classifier, point 9 is reweighted

Problem 4

The given plot of points can be represented as

X	Y	Class
1	1	-
2	2	-
1	3	-
2	6	-
4	1	-
4	4	-
5	1	-
8	1	-
9	1	-
2	8	+
5	9	+
6	8	+
6	5	+
7	3	+
7	7	+
8	7	+
8	9	+
9	4	+
9	6	+

The given test point is 5,4

using Euclidean distance, the distance of each point from the given test point is

1. $P_1(1,1)$ and $P(5,4)$

$\therefore d_1 = \sqrt{(5-1)^2 + (4-1)^2} = 5$

2. $P_2(2,2)$ and $P(5,4)$

$d_2 = \sqrt{(5-2)^2 + (4-2)^2} = \sqrt{13} = 3.605$

3. $P_3(1,3)$ and $P(5,4)$

$d_3 = \sqrt{(5-1)^2 + (4-3)^2} = \sqrt{17} = 4.123$

4. $P_4(2,6)$ and $P(5,4)$

$d_4 = \sqrt{(5-2)^2 + (4-6)^2} = \sqrt{13} = 3.605$

5. $P_5(4,1)$ and $P(5,4)$

$d_5 = \sqrt{(5-4)^2 + (4-1)^2} = \sqrt{10} = 3.162$

6. $P_6(4,4)$ and $P(5,4)$

$d_6 = \sqrt{(5-4)^2 + (4-4)^2} = 1$

7. $P_7(5,1)$ and $P(5,4)$

$d_7 = \sqrt{(5-5)^2 + (4-1)^2} = 3$

8. $P_8(8,1)$ and $P(5,4)$

$d_8 = \sqrt{(5-8)^2 + (4-1)^2} = \sqrt{18} = 4.242$

9. $P_9(9,1)$ and $P(5,4)$

$d_9 = \sqrt{(5-9)^2 + (4-1)^2} = \sqrt{25} = 5$

10. $P_{10}(2,8)$ and $P(5,4)$

$d_{10} = \sqrt{(5-2)^2 + (4-8)^2} = \sqrt{25} = 5$

11. $P_{11}(5,9)$ and $P(5,4)$

$d_{11} = \sqrt{(5-5)^2 + (4-9)^2} = 5$

12. $P_{12}(6,8)$ and $P(5,4)$

$d_{12} = \sqrt{(5-6)^2 + (4-8)^2} = \sqrt{17} = 4.123$

13. $P_{13}(6,5)$ and $P(5,4)$

$d_{13} = \sqrt{(5-6)^2 + (4-5)^2} = \sqrt{2} = 1.414$

$P_{14}(7,3)$ and $P(5,4)$

$d_{14} = \sqrt{(5-7)^2 + (4-3)^2} = \sqrt{5} \approx 2.236$

$P_{15}(7,7)$ and $P(5,4)$

$d_{15} = \sqrt{(5-7)^2 + (4-7)^2} = \sqrt{13} \approx 3.605$

$P_{16}(8,7)$ and $P(5,4)$

$d_{16} = \sqrt{(8-5)^2 + (7-4)^2} = \sqrt{18} = 4.242$

$P_{17}(8,9)$ and $P(5,4)$

$d_{17} = \sqrt{(5-8)^2 + (4-9)^2} = \sqrt{34} = 5.830$

$P_{18}(9,4)$ and $P(5,4)$

$d_{18} = \sqrt{(5-9)^2 + (4-4)^2} = 4$

$P_{19}(9,6)$ and $P(5,4)$

$d_{19} = \sqrt{(5-9)^2 + (4-6)^2} = \sqrt{20} = 4.472$

The 5 nearest neighbours are $P_2(4,4)$, $P_{13}(6,5)$, $P_{14}(7,3)$ $P_{5}(4,1)$, $P_7(5,7)$

2 Using Manhattan distance, the distance of each point from the test point is

$P_1(1,1)$ and $P(5,4)$

$d_1 = |5-1| + |4-1| = 7$

$P_2(2,2)$ and $P(5,4)$

$d_2 = |5-2| + |4-2| = 5$

$P_3(1,3)$ and $P(5,4)$

$d_3 = |5-1| + |4-3| = 4$

$P_4(2,6)$ and $P(5,4)$

$d_4 = |5-2| + |4-6| = 5$

$P_5(4,1)$ and $P(5,4)$

$d_5 = |5-4| + |4-1| = 4$

$P_6(4,4)$ and $P(5,4)$

$d_6 = |5-4| + |4-4| = 1$

$P_7(5,1)$ and $P(5,4)$

$d_7 = |5-5| + |4-1| = 3$

$P_8(8,1)$ and $P(5,4)$

$d_8 = |5-8| + |4-1| = 6$

$P_9(9,1)$ and $P(5,4)$

$d_9 = |5-9| + |4-1| = 7$

$P_{10}(2,8)$ and $P(5,4)$

$d_{10} = |5-2| + |4-8| = 7$

$P_{11}(5,9)$ and $P(5,4)$

$d_{11} = |5-5| + |4-9| = 5$

$P_{12}(6,8)$ and $P(5,4)$

$d_{12} = |5-6| + |4-8| = 5$

$P_{13}(6,5)$ and $P(5,4)$

$d_{13} = |5-6| + |4-5| = 2$

$P_{14}(7,3)$ and $P(5,4)$

$d_{14} = |5-7| + |4-3| = 3$

$P_{15}(7,7)$ and $P(5,4)$

$d_{15} = |5-7| + |4-7| = 5$

$P_{16}(8,7)$ and $P(5,4)$

$d_{16} = |5-8| + |4-7| = 6$

$P_{17}(8,9)$ and $P(5,4)$

$d_{17} = |5-8| + |4-9| = 8$

$P_{18}(9,4)$ and $P(5,4)$

$d_{18} = |5-9| + |4-4| = 4$

$P_{19}(9,6)$ and $P(5,4)$

$d_{19} = |5-9| + |4-6| = 6$

The 3 nearest neighbours are $P_6(4,4)$, $P_{13}(6,5)$, $P_7(8,3)$

weighted distance for $P_6 = \frac{1}{d_1^2} = \frac{1}{1}$

weighted distance for $P_3 = \frac{1}{d_2^2} = \frac{1}{4}$

weighted distance for $P_7 = \frac{1}{d_3^2} = \frac{1}{9}$

votes for ⊖ : 1

votes for ⊕ : $\frac{1}{4} + \frac{1}{9} = \frac{13}{36}$

- The test point will be classified as ⊖