UNIVERSITY OF WOLVERHAMPTON FACULTY OF SCIENCE AND ENGINEERING
7CS033 DATA MINING & INFORMATICS WORKSHOP # 6

Using the data point below , where the data points belong to three classes. We need to classify these using a decision tree classifier using one split.

X1	X2	Y
1	1	1
1	1	1
1	1	2
1	0	3
0	0	2
0	0	3

Information gain is a measure that is used to calculate effectiveness of an attribute in classifying a dataset. It quantifies the reduction in entropy or impurity after a dataset is split on an attribute.

Formula is given by :

$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

Pi = proportion of the points in class i:
From the table above we have 6 points from 3 classes :

- Class 1 : Pi = 2/6 = 0.33
- Class 2 : Pi = 2/6 = 0.33
- Class 3 : Pi = 2/6 = 0.33

Therefore :
H(s) = - (0.33Log2 0.33 + 0.33Log2 0.33 + 0.33Log2 0.33)

Root entropy **H(S) = 1.585**

Calculating the information gain for split on X1 :
We split as follows, based on values of 1 and 0.

- X1 = 1 : 4
- X1 = 0 : 2

Entropy of X1 = 1 :

H(X1= 1) = - (2/4Log2(2/4) + 1/4Log2(1/4) + 1/4Log212/4))

H(X1= 1)= 1.5

Entropy of X1 = 0 :

H(X1= 0) = - (1/2Log2(1/2) + 1/2Log2(1/2))

H(X1= 0)= 1

Weighted Entropy After Split:

Havg(X1) = 4/6***H(X1= 1) + 2/* H(X1= 0)** = 1.333

Finally information gain for X1 = H(S) - Havg(X1) = 1.585 **-** 1.333 = 0.252

Gain (X1) = 0.252

Calculating the information gain for split on X2 :
We split as follows, based on values of 1 and 0.

- X2 = 1 : 3
- X2 = 0 : 3

Entropy of X2 = 1 :

H(X2= 1) = - (2/3Log2(2/3) + 1/3Log2(1/3))

H(X2= 1) = 0.918

Entropy of X1 = 0 :

H(X2= 0) = - (1/3Log2(1/3) + 2/3Log2(2/3))

H(X2= 0) = 0.918

Weighted Entropy After Split:

Havg(X2) = 3/6***H(X2= 1) + 3/6* H(X2= 0)** = **0.918**

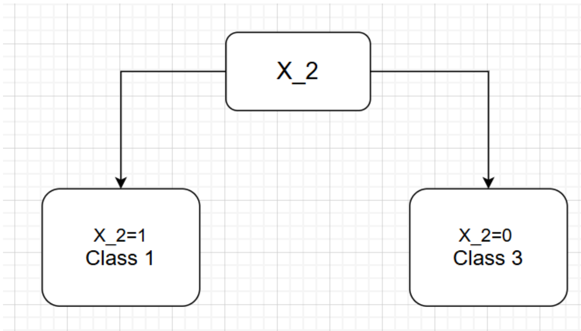
Finally information gain for X2 = H(S) - Havg(X2) = 1.585 **-** 0.918 = 0.667

Gain (X1) = 0.667

Compare the information gain for the two above we see that the split on X2 is higher than the information gain of x1 , hence reduces the uncertainty more effectively,

Diagram of decision tree and leaf labels :

Left branch : 2 points of Class 1, 1 point of Class 2 → Majority class = 1
Right branch : 1 point of Class 2, 2 points of Class 3 → Majority class = 3



K-Means clustering :

Use the K-means clustering to find two different clusters in the following sequence of three-dimensional points:
X=[(1,9,14),(2,18,23),(3,30,30),(4,21,9),(5,9,17),(6,25,32),(7,36,25),(8,10,12),(9,38,45),(10,1,2)]

Initial centroids randomly: (1,9,14) and (10,1,2).

C1 : (1,9,14).
C2 : (10,1,2).

Euclidean distance:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Point 1: (1,9,14) → C1=0

Point 2: (2,18,23) → C1=12.08, C2=25.16 → C1
Point 3: (3,30,30) → C1=23.64, C2=37.52 → C1
Point 4: (4,21,9) → C1=13.34, C2=20.88 → C1
Point 5: (5,9,17) → C1=5.0 , C2=17.26 → C1
Point 6: (6,25,32) → C1=22.36, C2=35.03 → C1
Point 7: (7,36,25) → C1=28.62 , C2=38.72 → C1
Point 8: (8,10,12) → C1=7.35 , C2=11.53 → C1
Point 9: (9,38,45) → C1=38.32, C2=50.12 → C1

Point 10: (10,1,2) → C1=17.15, C2=0 (same point) → C2

Clusters:

C1 : Points 1–9 -> (5,21.78,23), Taking the mean from 1-9.
C2 : Point 10 -> (10,1,2)

Iteration 2: Assign Clusters

Using new centroids C1=(5,21.78,23) , C2=(10,1,2) :

Point 1: C1=15.36, C2=17.15 → C1

Point 2: C1=5.10, C2=25.16 → C1

Point 3: C1=11.36, C2=37.52 → C1

Point 4: C1=14.00, C2=20.88→ C1

Point 5: C1=14.05, C2=17.26 → C1

Point 6: C1=9.92, C2=35.03 → C1

Point 7: C1=15.95 , C2=38.72→ C1

Point 8: C1=15.61, C2=11.53 → C2

Point 9: C1=24.54, C2=50.12→ C1

Point 10: C1=25.0, C2=0 → C2

Clusters:

C1: Points 1–7, 9 -> taking the mean -> (4.63,23.25,24.38)

C2: Points 8, 10 -> taking the mean -> (9,5.5,7)

Iteration 3: Assign Clusters

Point 1: C1=16.10 C2=11.05→ C2

Point 2: C1=6.09, C2=23.05 → C1

Point 3: C1=11.06, C2=35.95 → C1

Point 4: C1=15.23, C2=13.80→ C2

Point 5: C1=15.34, C2=12.18→ C2

Point 6: C1=9.62, C2=33.40 → C1

Point 7: C1=15.06, C2=36.37 → C1

Point 8: C1=17.31, C2=5.50 → C2

Point 9: C1=24.07, C2=47.81→ C1

Point 10: C1=25.43 , C2=6.58 → C2

Clusters:

- C1: Points 2, 3, 6, 7, 9 -> taking the mean (5.4,29.4,31)
- C2: Points 1, 4, 5, 8, 10 -> taking the mean (5.6,10,10.8)

Final Clusters:

- **Cluster 1:** (2,18,23) , (3,30,30), (6,25,32), (7,36,25), (9,38,45).
- **Cluster 2:** (1,9,14) , (4,21,9), (5,9,17), (8,10,12), (10,1,2).

Centroids: C1=(5.4,29.4,31), C2=(5.6,10,10.8)