# C4.5

Formulae

$$\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}$$

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_i|}{|D|} * \log_2 \frac{|D_j|}{|D|}$$

$$\text{SplitInfo}(D) = -p(\text{poor}) \log_2(p(\text{poor})) - p(\text{avg}) \log_2(p(\text{avg})) - p(\text{excallent}) \log_2(p(\text{excallent}))$$

$$= -\frac{7}{20} \log\left(\frac{7}{20}\right) - \frac{7}{20} \log\left(\frac{7}{20}\right) - \frac{6}{20} \log\left(\frac{6}{20}\right)$$

$$= 1.581$$

Entropy (D) = 1.581

Academics

| | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| po fail | 2 | 0 | 1 | 3 | 0.9182 |
| pass | 5 | 6 | 0 | 11 | 0.994 |
| distinction | 0 | 1 | 5 | 6 | 0.65 |

$$\text{entropy}(\text{academics} = \text{'fail'}) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$$

$$= 0.918$$

$$E(\text{academics} = \text{'pass'}) = -\frac{5}{11}\log_2\left(\frac{5}{11}\right) - \frac{6}{11}\log_2\left(\frac{6}{11}\right)$$

$$= 0.994$$

$$E(\text{academics} = \text{'distinction'}) = -\frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{5}{6}\log_2\left(\frac{5}{6}\right)$$

$$= 0.65$$

$$I(\text{academics}) = \left[\left(\frac{2+0+1}{20}\right) * 0.918\right] + \left[\left(\frac{11}{20}\right) * 0.994\right] + \left[\frac{6}{20} * 0.65\right]$$

$$= 0.8794$$

$$\text{Gain} = \text{Entropy}(D) - I(A)$$

$$= 1.581 - 0.8794 \quad = 0.7016$$

$$\text{Splitinfo}(\text{'academics'}) = \frac{-3}{20}\log_2\left(\frac{3}{20}\right) - \frac{11}{20}\log_2\left(\frac{11}{20}\right) - \frac{6}{20}\log_2\left(\frac{6}{20}\right)$$

$$= 1.406$$

$$\text{gainratio}(\text{academics}) = \frac{\text{gain}}{\text{splitinfo}} = \frac{0.7016}{1.406} = 0.499$$

**Speaking**

| | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| hesitant | 6 | 2 | 2 | 10 | 1.371 |
| fair | 1 | 5 | 0 | 6 | 0.65 |
| fluent | 0 | 0 | 4 | 4 | 0 |

$$E(\text{speaking} = \text{'hesitant'}) = \frac{-6}{10}\log_2\left(\frac{6}{10}\right) - \frac{2}{10}\log_2\left(\frac{2}{10}\right) - \frac{2}{10}\log_2\left(\frac{2}{10}\right) = 1.371$$

$$E(\text{speaking} = \text{'fair'}) = \frac{-1}{6}\log_2\left(\frac{1}{6}\right) - \frac{5}{6}\log_2\left(\frac{5}{6}\right) - 0$$

$$= 0.65$$

$$E(\text{speaking} = \text{'fluent'}) = 0$$

~~Gain (speaki~~ $I(\text{speaking}) = \left(\frac{1.371}{20} * 10\right) + \left(\frac{6}{20} * 0.65\right)$

$$= 0.8805$$

$= E(S) - I(\text{Speaking})$

Gain (speaking) $= 1.581 - 0.8805$

$= 0.7005$

Splitinfo(~~A~~ speaking) $= -\frac{10}{20}\log_2\left(\frac{10}{20}\right) - \frac{6}{20}\log_2\left(\frac{6}{20}\right) - \frac{4}{20}\log_2\left(\frac{4}{20}\right)$

$= 1.485$

Gain ratio (speaking) $= \frac{\text{Gain}}{\text{SplitInfo}} = \frac{0.7005}{1.485} = .47171$

**Creative**

| | poor | avg | exallent | total | entropy |
|---|---|---|---|---|---|
| low | 4 | 1 | 0 | 5 | .722 |
| medium | 3 | 4 | 3 | 10 | 1.571 |
| high | 0 | 2 | 3 | 5 | 0.971 |

E(creative = 'low') $= -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) = 0.722$

E(creative = 'medium') $= -\frac{3}{10}\log_2\left(\frac{3}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right) - \frac{3}{10}\log_2\left(\frac{3}{10}\right)$

$= 1.571$

E(creative = 'high') $= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = .971$

I(creative) $= \left(\frac{5}{20} * .722\right) + \left(\frac{10}{20} * 1.571\right) + \left(\frac{5}{20} * .971\right)$

$= 1.208$

gain(creative) = E(creative) − I (creative)

$= 1.581 - 1.208 = 0.37225$

~~gainratio~~ splitinfo (creative) $= -\frac{5}{20}\log_2\left(\frac{5}{20}\right) - \frac{10}{20}\log_2\left(\frac{10}{20}\right) - \frac{5}{20}\log_2\left(\frac{5}{20}\right)$

$= 1.5$

gain ratio (creative) $= \frac{0.37225}{1.5} = 0.24816$

| Sports | poor | avg | excallent | total | entropy |
|---|---|---|---|---|---|
| bad | 7 | 3 | 0 | 10 | 0.881 |
| good | 0 | 3 | 3 | 6 | 1 |
| v. good | 0 | 1 | 3 | 4 | 0.811 |

$E(\text{sports='bad'}) = -\frac{7}{10}\log_2\left(\frac{7}{10}\right) - \frac{3}{10}\log_2\left(\frac{3}{10}\right) - 0$

$= 0.881$

$E(\text{sports= 'good'}) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) - 0 = 1$

$E(\text{sports= 'very good'}) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) - 0$

$= 0.811$

$I(\text{sports}) = \left[\frac{10}{20} * 0.881\right] + \left[\frac{6}{20} * 1\right] + \left[\frac{4}{20} * 0.811\right]$

$= .9027$

$Gain(\text{sports}) = E(S) - I(\text{sports})$

$= 1.581 - 0.9027$

$= 0.6783$

$\text{splitinfo}(\text{sports}) = -\frac{10}{20}\log\left(\frac{10}{20}\right) - \frac{6}{20}\log\left(\frac{6}{20}\right) - \frac{4}{20}\log\left(\frac{4}{20}\right)$

$= 1.485$

$\text{gain ratio}(\text{sports}) = \frac{\text{gain(sport)}}{\text{splitinfo (sport)}} = \frac{.6783}{1.485} = .456$

| Att. | gain ratio |
|---|---|
| Academics | .499 ✓ |
| Speaking | .4717 |
| Creative | .24816 |
| Sports | .456 |

'Academics' attribute has highest gain ratio hence 'Academics' is the root node.

Academics

fail ← ↓ pass → distinction

Academics → ~~fail~~ distinction

data points = 6

$E(s)_{\text{distinction}} = -\frac{5}{6}\log_2\left(\frac{5}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right) = 0.65$

Speaking

| | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| hesitant | 0 | 1 | 1 | 2 | 1 |
| fair | 0 | 0 | 0 | 0 | 0 |
| fluent | 0 | 0 | 4 | 4 | 0 |

$E(\text{speaking} = \text{'hesitant'}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

$E(\text{speaking} = \text{'fair'}) = 0$

$E(\text{speaking} = \text{'fluent'}) = 0$

$I(\text{speaking}) = \frac{2}{6} * 1 = 0.333$

$\text{Gain (speaking)} = E(S_{\text{distinction}}) - I = 0.65 - 0.33$
$= 0.32$

$\text{splitinfo} = -\frac{2}{6}\log\left(\frac{2}{6}\right) - \frac{4}{6}\log\left(\frac{4}{6}\right) = 0.918$

$\text{gain ratio( speaking)} = \frac{0.32}{0.918} = 0.3485$

Creative

| | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| low | 0 | 1 | 0 | 1 | 0 |
| medium | 0 | 0 | 3 | 3 | 0 |
| high | 0 | 0 | 2 | 2 | 0 |

$E(\text{'creative} = \text{'low'}) = 0$

$E(\text{creative} = \text{'med'}) = 0$ $\quad E(\text{creative} = \text{'high'}) = 0$

$I(\text{creative}) = 0$

$\text{Gain (creative)} = 0.65 - 0 = 0.65$

$\text{splitinfo(creative)} = -\frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 1.459$

Gain ratio $= \frac{0.65}{1.459} = 0.4455$

| Sports | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| bad | 0 | 0 | 0 | 0 | 0 |
| good | 0 | 1 | 3 | 4 | 0.811 |
| very good | 0 | 0 | 2 | 2 | 0 |

$E(\overset{\text{sports=}}{\text{bad}}) = 0$

$E(\text{sport} = \text{'good'}) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.811$

$E(\text{sport} = \text{'very good'}) = 0$

$I(\text{sport}) = \frac{4}{2} * 0.811 = 0.5406$

$\text{Gain}(\text{sport}) = \frac{4}{8}\ E(s) - I(\text{sport}) = 0.65 - 0.5406 = 0.1094$

Splitinfo $= 1.028$ Gain ratio $= 0.10642$

| Attribute | Gain Ratio | |
|---|---|---|
| Speaking | 0.348 | |
| Creative | 0.445 | ✓ |
| Sports | 0.106 | |

Choosing creative as the next level node because it has highest gain ratio.

$$\text{Academics} \rightarrow \text{distinction} \rightarrow \text{creative} \begin{cases} \rightarrow \text{low} \rightarrow \text{avg} \\ \rightarrow \text{med} \rightarrow \text{excellent} \\ \rightarrow \text{high} \rightarrow \text{excellent} \end{cases}$$

Academics → pass

data points $= 11$ $E(S_{\text{pass}}) = 0.994$

| Speaking | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| hesitant | 4 | 1 | 0 | 5 | 0.722 |
| fair | 1 | 5 | 0 | 6 | 0.65 |
| fluent | 0 | 0 | 0 | 0 | 0 |

$I = \left[\left(\frac{5}{11}\right) * 0.722\right] + \left[\left(\frac{6}{11}\right) * 0.65\right] = 0,6827$

Gain = 0,3113    Splitinfo = 0.994    Gainratio = 0,313

| Creative | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| low | 3 | 0 | 0 | 3 | 0 |
| med | 2 | 4 | 0 | 6 | 0,918 |
| high | 0 | 2 | 0 | 2 | 0 |

$I = \left[\frac{6}{11} * .918\right] = 0,5$    Gain = .494

Splitinfo = 1,435    Gainratio = .344

| Sports | poor | avg | exallent | total | entropy |
|---|---|---|---|---|---|
| bad | 5 | 3 | 0 | 8 | 0,954 |
| good | 0 | 2 | 0 | 2 | 0 |
| very good | 0 | 1 | 0 | 1 | 0 |

$I = \left(\frac{8}{11} * .954\right) = .6938$

Gain = .994 − .6938 = 0.3

Splitinfo = 1,096    Gain ratio = 0,273

| Attribute | gainratio |
|---|---|
| Speaking | .313 |
| Creative | .344 ✓ |
| Sports | .273 |

Academics → pass → creative → med → ?

datapoints = 6    entropy = 0,918

| Speaking | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| hesitant | 2 | 0 | 0 | 2 | 0 |
| fair | 0 | 4 | 0 | 4 | 0 |
| fluent | 0 | 0 | 0 | 0 | 0 |

I=0 gain= 0.918

splitinfo = 0.918 gain ratio =1

| Sports | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| bad | 2 | 2 | 0 | 4 | 1 |
| good | 0 | 1 | 0 | 1 | 0 |
| v.good | 0 | 1 | 0 | 1 | 0 |

I = .666 gain = .2514 splitinfo = 1.252

gain ratio = 0.2

| Attribute | gain ratio |
|---|---|
| Speaking | 1 |
| Sports | 0.2 |

Thus next node is Speaking.

Academics → fail data point = 3

E = 0.918

| Speaking | poor | avg | exellent | total | entropy |
|---|---|---|---|---|---|
| hesitant | 2 | 0 | 1 | 3 | .918 |
| fair | 0 | 0 | 0 | 0 | 0 |
| fluent | 0 | 0 | 0 | 0 | 0 |

I = .918 gain = 0

splitinfo =0 gainratio = 0

| Creative | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| low | 1 | 0 | 0 | 1 | 0 |
| med | 1 | 0 | 0 | 1 | 0 |
| high | 0 | 0 | 1 | 1 | 0 |

I=0

gain = .918 splitinfo = 1.585 gainratio = .579

| Sports | poor | avg | excellent | total | entropy |
|---|---|---|---|---|---|
| bad | 2 | 0 | 0 | 2 | 0 |
| good | 0 | 0 | 0 | 0 | 0 |
| very good | 0 | 0 | 1 | 1 | 0 |

I=0

gain = .918 splitinfo = .918 gainratio = 1

| Attribute | gain ratio | |
|---|---|---|
| speaking | 0 | |
| Creative | .579 | |
| Sports | 1 | ✓ |

In this case C4.5 proves to be better than ID3 as when the same scenario occured in ID3 then the gain was covered which was same for 'Creative' & 'Sports' attribute. But here 'sports' has higher gain ratio thus becomes the next level node of the decision tree.