Decision Tree Worksheet

Class label: "Passed Exam"

Attributes: "Passed all assignments", "GPA", "Language"

(1) Impurity of Parent w/ Gini Index

$$I(t) = 1 - \sum_i p_i^2 = 1 - \left[\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right] \approx 0.489$$

(2) Impurity of child nodes when splitting on "passed all assignments"

When "Passed all assignments" == No, v_1

$$I(v_1) = 1 - \sum_i p_i^2 = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.445$$

$$N(v_1) = 3$$

When "passed all assignments == Yes, v_2

$$I(v_2) = 1 - \sum_i p_i^2 = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$$

$$N(v_2) = 4$$

Gain is

$$I(t) - \sum_i \frac{N(v_j)}{N} I(v_j)$$

$$= 0.489 - \left[\frac{3}{7}(0.445) + \frac{4}{7}(0.5)\right] = 0.012$$

(3) Impurity of child nodes when splitting on language

When Language == Python, v_1

$$I(v_1) = 1 - \sum_i p_i^2 = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.445$$

$$N(v_1) = 3$$

Language == C++ , v_2

$$I(v_2) = 1 - \sum_i p_i^2 = 1 - \left[\left(\frac{2}{2}\right)^2 + 0\right] = 0$$

$$N(v_2) = 2$$

Language == Java

$$I(v_3) = 1 - \sum_i p_i^2 = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$$

$$N(v_3) = 2$$

$$\rightarrow \text{Gain} = I(t) - \sum_j \frac{N(v_j)}{N} I(v_j)$$

$$= 0.489 - \left[\frac{3}{7}(0.445) + \frac{2}{7}\cdot 0 + \frac{2}{7}0.5\right] = 0.156$$

(4) Gain from using GPA

Step 1 sort GPA

GPA	Passed Exam
2.0	No
2.5	No
3.1	yes
3.2	yes
3.3	yes
3.5	yes
3.9	No

Only need to consider split points where there is a change in class

Candidate split point $\frac{2.5+3.1}{2} = 2.8$

Candidate split point $\frac{3.5+3.9}{2} = 3.7$

Calculate the weighted Impurity with two possible binary splits (for purposes of example – could consider more)

GPA ≤ 2.8

$I(v_1) = 0$

$N(v_1) = 2$

GPA > 2.8

$I(v_2) = 1 - \sum_i p_i^2 = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0.32$

$N(v_2) = 5$

Weighted Gini Index $= \frac{2}{7}\cdot 0 + \frac{5}{7}(0.32) = 0.228$

GPA ≤ 3.7	GPA > 3.7
$I(v_1) = 1 - \left[\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right] = 0.445$	$I(v_2) = 1 - 1 = 0$
$N(v_1) = 6$	$N(v_2) = 1$

Weighted Gini Index $= \frac{6}{7}(0.445) + 0.1 = 0.381$

Splitting on GPA = 2.8 is better, lowers impurity

Gain is $0.489 - 0.228 = 0.261$

(3) Split Info for all 3 ways

$$\text{Split Info} = -\sum_j \frac{N(v_j)}{N} \log_2\left(\frac{N(v_j)}{N}\right)$$

Passed all assignments

$$\text{Split Info} = -\left[\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$$

$$= 0.985$$

$$\text{Gain Ratio is } \frac{\text{Gain}}{\text{Split Info}} = \frac{0.012}{0.985} \approx 0.012$$

Language

$$\text{Split Info} = -\left[\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right]$$

$$= 1.556$$

Gain Ratio $0.156 / 1.556 = 0.1$

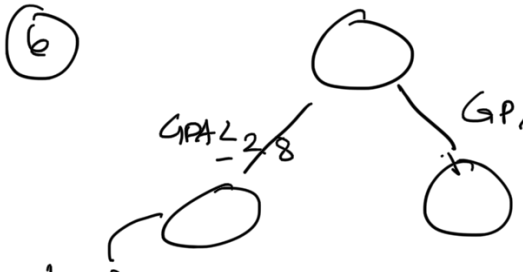
GPA

Split Info = $-\left[\frac{7}{7}\log_2\left(\frac{7}{7}\right) + \frac{7}{7}\log_2\left(\frac{7}{7}\right)\right] = 0.86$.

Gain Ratio = $0.261 / 0.863 = 0.302$

$\Rightarrow$ GPA is the first Attribute to split on

(6)



Leaf Node

Impurity = 0

Impurity = 0.32 as calculated above