# Practice Problem 1

Ian Chen

February 1, 2024

**Problem 0** — What is the class label? What are the attributes?

**Answer**
*Class label- Passed exam*
*Attributes- Passed all assignments, GPA, Language*

**Problem 1** — Start by calculating the impurity of the parent. Use entropy as the measure of impurity.

**Answer**
*Entropy* $= \sum_{i=1}^{n} -p_i \log_2 p_i$
*There are 3 NOs and 4 YESs*
$p_{NO} = \frac{3}{7}$
$p_{YES} = \frac{4}{7}$
*Entropy* $= -p_{NO} \log_2(p_{NO}) - p_{YES} \log_2(p_{YES})$
$= -\frac{3}{7} \log_2(\frac{3}{7}) - \frac{4}{7} \log_2(\frac{4}{7})$
= *0.985228136*

**Problem 2** — Next, calculate the information gain of splitting on 'Passed All Assignments'. The gain of splitting on 'Passed All Assignments' will be the entropy of the parent minus the entropy of making this split. (We want to know how much the impurity decreases by making this split.)

**Answer**

| | *Didn't Pass All Assignments* | *Passed All Assignments* |
|---|---|---|
| *NO* | *1* | *2* |
| *YES* | *2* | *2* |

*Left branch:* $p_{NO} = \frac{1}{3}$, $p_{YES} = \frac{2}{3}$
*Entropy* $= -\frac{1}{3} log_2(\frac{1}{3}) - \frac{2}{3} log_2(\frac{2}{3})$
= *0.9182958341*
*Right branch:* $p_{NO} = \frac{2}{4}$, $p_{YES} = \frac{2}{4}$
*Entropy* $= -\frac{1}{2} log_2(\frac{1}{2}) - \frac{1}{2} log_2(\frac{1}{2})$
= *1*
*Information Gain = Entropy(parent) -* $(\sum_{i=1}^{k} \frac{n_i}{n} Impurity(i))$
= *0.985228136* - $(\frac{3}{7}0.9182958341 + \frac{4}{7}1)$
= *0.0202442071*

**Problem 3** — Next, calculate the information gain of splitting on 'Language'. The gain of splitting on

'Language' will be the entropy of the parent minus the entropy of making this split. (We want to know how much the impurity decreases by making this split.)

**Answer**

| | *Python* | *Java* | *C++* |
|---|---|---|---|
| *NO* | *2* | *1* | *0* |
| *YES* | *1* | *1* | *2* |

*Left branch:* $p_{NO} = \frac{2}{3}$, $p_{YES} = \frac{1}{3}$

*Entropy* $= -\frac{2}{3} log_2(\frac{2}{3}) - \frac{1}{3} log_2(\frac{1}{3})$

$= 0.9182958341$

*Middle branch:* $p_{NO} = \frac{1}{2}$, $p_{YES} = \frac{1}{2}$

*Entropy* $= -\frac{1}{2} log_2(\frac{1}{2}) - \frac{1}{2} log_2(\frac{1}{2})$

$= 1$

*Right branch:* $p_{NO} = \frac{0}{2}$, $p_{YES} = \frac{2}{2}$

*Entropy* $= -\frac{0}{2} log_2(\frac{0}{2}) - \frac{2}{2} log_2(\frac{2}{2})$

$= 0$

*Information Gain* $= 0.985228136$ - $(\frac{3}{7} \cdot 0.9182958341 + \frac{2}{7} \cdot 1 + \frac{2}{7} \cdot 0)$

$= 0.3059584928$

**Problem 4** — Next, calculate the information gain of splitting on 'GPA'. Because GPA is a continuous attribute, we need to try different candidate threshold values. Determine all of the candidate thresholds, then calculate the information gain for each of them.

**Answer**

**Problem 5** — Impurity metrics(like Gini and entropy) favor attributes with more values. Because 'Language' can be split 3 ways, but 'Passed All Assignments' and 'GPA' are only split 2 ways, we should use gain ratio to compare the attributes, rather than just gain. Calculate the split info for each of the three attributes. With that, calculate the gain ratio for each of the three attributes.

**Answer**