Practice Problem on Decision Trees

Given the weather conditions, we want to predict if a person is going for a run or not. The data that we have collected are the following:

	Sample	**Features**		**Outcome**
		Outlook	**Wind**	**Run**
Train	**S1**	Sunny	Weak	No
	S2	Sunny	Strong	No
	S3	Overcast	Weak	Yes
	S4	Rain	Weak	Yes
	S5	Rain	Weak	Yes
	S6	Rain	Strong	No
	S7	Overcast	Strong	Yes

Based on the above data, we will build a decision tree using the entropy splitting criterion. The input features are **Outlook** and **Wind**, while the outcome variable is **Run**.

(a) Compute the entropy splitting criterion of the outcome **Run** conditioned on the **Outlook** and **Wind** features. Which feature will be used as the splitting attribute in root of the tree? Show all your calculations.
Note: You **do not** need to perform arithmetic calculations for logarithms, e.g. if one of your equations contains $\log(\frac{1}{3})$, you can leave it like that and still solve the problem.

We will compute the partial information entropy for each feature value. Then we will obtain the information entropy of each feature using a weighted mean, where the weights correspond to the probability of each value. In the following calculations, by convention, we ignore that $log0 \to \infty$, and assume that the corresponding terms contribute with zero.

$$H(Run|Outlook = Sunny) = -\left[\tfrac{2}{0+2}log\tfrac{2}{0+2} + \tfrac{0}{0+2}log\tfrac{0}{0+2}\right] = 0$$

$$H(Run|Outlook = Overcast) = -\left[\tfrac{0}{2+0}log\tfrac{0}{2+0} + \tfrac{2}{2+0}log\tfrac{2}{2+0}\right] = 0$$

$$H(Run|Outlook = Rain) = -\left[\tfrac{1}{1+2}log\tfrac{1}{1+2} + \tfrac{2}{1+2}log\tfrac{2}{1+2}\right] = -\tfrac{1}{3}log\tfrac{1}{3} - \tfrac{2}{3}log\tfrac{2}{3}$$

$$\begin{aligned} H(Run|Outlook) &= \frac{2}{7} \times H(Run|Outlook = Sunny) + \frac{2}{7} \times H(Run|Outlook = Overcast) \\ &+ \frac{3}{7} \times H(Run|Outlook = Rain) \\ &= -\frac{3}{7}\left(\frac{1}{3}log\frac{1}{3} + \frac{2}{3}log\frac{2}{3}\right) = \frac{3}{7}\left(\frac{1}{3}log3 + \frac{2}{3}log\frac{3}{2}\right) \end{aligned}$$

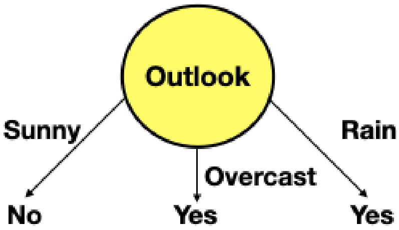
$$H(Run|Wind = Weak) = -\left[\tfrac{1}{1+3}log\tfrac{1}{1+3} + \tfrac{3}{1+3}log\tfrac{3}{1+3}\right] = -\tfrac{1}{4}log\tfrac{1}{4} - \tfrac{3}{4}log\tfrac{3}{4}$$

$$H(Run|Wind = Strong) = -\left[\tfrac{2}{2+1}log\tfrac{2}{2+1} + \tfrac{1}{2+1}log\tfrac{1}{2+1}\right] = -\tfrac{2}{3}log\tfrac{2}{3} - \tfrac{1}{3}log\tfrac{1}{3}$$

$$\begin{aligned}
H(Run|Wind) &= \frac{4}{7} \times H(Run|Wind = Weak) + \frac{3}{7} \times H(Run|Wind = Strong) \\
&= -\frac{4}{7}\left(\frac{1}{4}log\frac{1}{4} + \frac{3}{4}log\frac{3}{4}\right) - \frac{3}{7}\left(\frac{2}{3}log\frac{2}{3} + \frac{1}{3}log\frac{1}{3}\right) \\
&= \frac{4}{7}\left(\frac{1}{4}log4 + \frac{3}{4}log\frac{4}{3}\right) + \frac{3}{7}\left(\frac{2}{3}log\frac{3}{2} + \frac{1}{3}log3\right)
\end{aligned}$$

We observe that $H(Run|Outlook) < H(Run|Wind)$, therefore **Outlook** will be the splitting criterion for the root node.

(b) Create the decision tree using only one node, i.e., the tree will only have the root. Please show the **splitting criterion of the node**, as well as **the decisions from each possible outcome of the corresponding criterion**. Please **describe how the decisions were made**.



All samples with $Outlook = Sunny$ correspond to $Run = No$.

All samples with $Outlook = Overcast$ correspond to $Run = Yes$.

The majority of samples (i.e., 2 out of 3) with $Outlook = Rain$ correspond to $Run = Yes$, therefore we will be taking this decision if we make this as a terminal node. If we select not to make this as a terminal node, then we will keep expanding the tree.

(c) Which of the training samples will be classified correctly only using the above tree and which not?

Classified correctly: $S1$, $S2$, $S3$, $S4$, $S5$, $S7$

Classified incorrectly: $S6$

(d) Expand the tree until all samples are correctly classified.

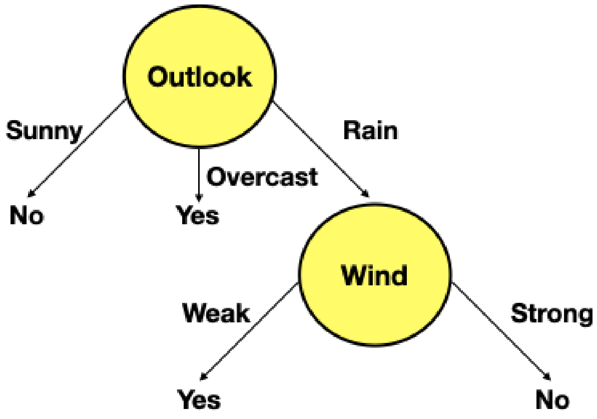
We will compute the information entropy for both features only for the samples that have reached the node $Outlook = Rain$ (i.e., $S4$, $S5$, $S6$).

$H(Run|Outlook) = H(Run|Outlook = Rain) = -\frac{1}{3}log\frac{1}{3} - \frac{2}{3}log\frac{2}{3}$

$H(Run|Wind = Weak) = 0$, since the outcome is Yes for both $S4$ and $S5$, for which $Wind = Weak$

$H(Run|Wind = Strong) = 0$, since the outcome is No for $S6$, for which $Wind = Strong$

Therefore $H(Run|Wind) < H(Run|Outlook)$, so we will pick feature **Wind** in this node, and the tree will looks as follows:


Outlook
Sunny
Overcast
Rain
No
Yes
Wind
Weak
Strong
Yes
No