Decision Trees Example with Entropy

Example from Sadawi/Sayad video lesson at https://youtu.be/O__7lAqni7A.

	FEATURES				TARGET
#	**Outlook**	**Temp**	**Humidity**	**Windy**	**Play Golf**
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

Initial entropy calculation, prior to splitting on any features:

Play Golf	
Yes	No
9	5

Here's the entropy formula:

$$E = \sum_{i=1}^{C} -p_i * log_2(p_i)$$

Where:

C = # target values, in this dataset C = 2 (Yes or No)

i = index [1...C] of each target value, in this dataset i = 1 or 2.

p_i = the frequency fraction of data cases for target value (i)

So the summation evaluates to:

$$\begin{aligned} E &= -(\tfrac{9}{14})\, log_2(\tfrac{9}{14}) - (\tfrac{5}{14})\, log_2(\tfrac{5}{14}) \\ &= -(.643)\, log_2(.643) - (.357)\, log_2(.357) = 0.940 \end{aligned}$$

If we first split on the feature Outlook, we get the following:

		Play Golf		
		Yes	**No**	**Total**
Outlook	**Sunny**	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
			All	14

Each of the three splits, Sunny, Overcast and Rainy, has a corresponding group of cases, with an entropy. These three entropy calculations must be combined, weighting for the size, to generate a total entropy.

$$E_{total} = \sum_{c \in X} P(c) * E(c)$$

Where:

C = feature value

X = set of feature values, here Sunny, Overcast or Rainy

$P(c)$ = frequency fraction of cases with feature value = c

$$\begin{aligned} E_{total} &= P(Sunny) * E(Sunny) \\ &+ P(Overcast) * E(Overcast) \\ &+ P(Rainy) * E(Rainy) \end{aligned}$$

$$= \ (\tfrac{5}{14}) * 0.971 \ + \ (\tfrac{4}{14}) * 0.0 \ + \ (\tfrac{5}{14}) * 0.971 \ = \ 0.693$$

Since

$$E(Sunny) \ = \ -(\tfrac{3}{5})\, log_2(\tfrac{3}{5}) - (\tfrac{2}{5})\, log_2(\tfrac{2}{5}) \ = \ 0.971$$
$$E(Overcast) \ = \ -(\tfrac{4}{4})\, log_2(\tfrac{4}{4}) - (\tfrac{0}{4})\, log_2(\tfrac{0}{4}) \ = \ 0.0$$
$$E(Rainy) \ = \ -(\tfrac{2}{5})\, log_2(\tfrac{2}{5}) - (\tfrac{3}{5})\, log_2(\tfrac{3}{5}) \ = \ 0.971$$

Information Gain = Entropy Decrease = -(E_{final} - $E_{initial}$)

So for this split,

Information Gain = - (.693 - .940) = .247

Constructing a **decision tree** is all about finding criteria to **maximize information gain** at each split. In other words, the goal is to reduce **mis-classification error**, and increase the **homogeneity** or **purity** of each branch.