Decision Tree

Definition 0.1. The basic decision tree learning algorithmn uses entropy to measure homogeneity of data

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2(p_i)$$

with $p_i = \frac{p_i}{\sum_i^c p_i}$ when $1 \leq i \leq c = |S|$ holds.

Definition 0.2. The information gain function defined as

$$Gain(S, A) = Entropy(S) - \sum_{v_i \in Values(A)}^{c} \frac{|S_v|}{|S|} Entropy(S_v)$$

Example 0.3. The table below will be use for the following excercises and examples.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Solution 1.5. The formula is $Entropy(S)$ derived as

$$-[p_1 log_2(p_1) + p_2 log_2(p_2)] = -\frac{9}{14} log_2(\frac{9}{14}) - \frac{5}{14} log_2(\frac{5}{14}) = 0.94028$$

with $p_1 = \frac{p_1}{p_1+p_2} = \frac{9}{14}$ and $p_2 = \frac{p_2}{p_1+p_2} = \frac{5}{14}$ where $S = [9+, 5-]$

Gain(S_{sunny},O) with O as outcast

Solution 1.6. The formula is $Entropy(S_{sunny})$ derived as

$$-[p_1 log_2(p_1) + p_2 log_2(p_2)]$$

$$-\frac{2}{5}log_2(\frac{2}{5}) - \frac{3}{5}log_2(\frac{3}{5})$$

$$0.97095$$

with $p_1 = \frac{p_1}{p_1+p_2} = \frac{2}{5}$ and $p_2 = \frac{p_2}{p_1+p_2} = \frac{3}{5}$ where $S_{sunny} = [2+, 3-]$

Note 1.7. $\lim_{x \to 0^+} xlog(x) = 0$ and $\lim_{x \to 0^-} xlog(x) = 0$.

Solution 1.7. The formula is $Entropy(S_{overcast})$ derived as

$$-[p_1 log_2(p_1) + p_2 log_2(p_2)]$$

$$-\frac{4}{4}log_2(\frac{4}{4}) - \frac{0}{4}log_2(\frac{0}{4})$$

$$0$$

with $p_1 = \frac{p_1}{p_1+p_2} = \frac{4}{4}$ and $p_2 = \frac{p_2}{p_1+p_2} = \frac{0}{4}$ where $S_{overcast} = [4+, 0-]$

Solution 1.8. The formula is $Entropy(S_{rain})$ derived as

$$-[p_1 log_2(p_1) + p_2 log_2(p_2)]$$

$$-\frac{3}{5}log_2(\frac{3}{5}) - \frac{2}{5}log_2(\frac{2}{5})$$

$$0.97095$$

with $p_1 = \frac{p_1}{p_1+p_2} = \frac{3}{5}$ and $p_2 = \frac{p_2}{p_1+p_2} = \frac{2}{5}$ where $S_{rain} = [3+, 2-]$

Solution 1.9. The formula to compute Gain(S,O) is as followed with Values(O) with O as outcast

$$Entropy(S) - \frac{|S_{sunny}|}{|S|}Entropy(S_{sunny}) - \frac{|S_{Overcast}|}{|S|}Entropy(S_{Overcast}) - \frac{|S_{rain}|}{|S|}Entropy(S_{rain})$$

$$0.94 - \frac{5}{14}0.97 - \frac{4}{14}0 - \frac{5}{14}0.97$$

$$0.247$$

Gain(S,T) with T as temperature

Note 2.1. $log_{1/2} = log_2(1) - log_2(2) = -1$

Solution 2.1. The formula is $Entropy(S_{Hot})$ derived as

$$-[p_1 log_2(p_1) + p_2 log_2(p_2)]$$

$$-\frac{2}{4}log_2(\frac{2}{4}) - \frac{2}{4}log_2(\frac{2}{4})$$

$$-\frac{1}{2}(-1) + -\frac{1}{2}(-1)$$

$$1$$

with $p_1 = \frac{p_1}{p_1+p_2} = \frac{2}{4}$ and $p_2 = \frac{p_2}{p_1+p_2} = \frac{2}{4}$ where $S = [2+, 2-]$

Solution 2.2. The formula is $Entropy(S_{Mild})$ derived as

$$-[p_1 log_2(p_1) + p_2 log_2(p_2)]$$

$$-\frac{4}{6}log_2(\frac{4}{6}) - \frac{2}{6}log_2(\frac{2}{6})$$

$$.918$$

with $p_1 = \frac{p_1}{p_1+p_2} = \frac{4}{6}$ and $p_2 = \frac{p_2}{p_1+p_2} = \frac{2}{6}$ where $S = [4+, 2-]$

Solution 2.3. The formula is $Entropy(S_{Cool})$ derived as

$$-[p_1 log_2(p_1) + p_2 log_2(p_2)]$$

$$-\frac{3}{4}log_2(\frac{3}{4}) - \frac{1}{4}log_2(\frac{1}{4})$$

$$0.811$$

with $p_1 = \frac{p_1}{p_1+p_2} = \frac{3}{4}$ and $p_2 = \frac{p_2}{p_1+p_2} = \frac{1}{4}$ where $S = [3+, 1-]$

Solution 2.4. The formula to compute Gain(S,T) is as followed with Values(T) with T as temperature

$$Entropy(S) - \frac{|S_{Hot}|}{|S|}Entropy(S_{Hot}) - \frac{|S_{Mild}|}{|S|}Entropy(S_{Mild}) - \frac{|S_{Cool}|}{|S|}Entropy(S_{Cool})$$

$$0.94 - \frac{4}{14}1 - \frac{6}{14}.918 - \frac{4}{14}.811$$

$$0.029$$

Gain(S_{Rain}, R_w) with R as rain and subscript W as wind

Solution 3.1. $Entropy(R_{weak})$ and $Entropy(R_{strong})$ both are 0 with $R_{weak} = [3+, 0-]$ and $R_{weak} = [0+, 2-]$, see solution 1.7. The $Entropy(S_{rain})$ is 0.97, see solution 1.8. Final, the Gain(S_{weak}, R_w) is computed as

$$Entropy(S_{Rain}) - \frac{|R_{Weak}|}{|S_{Rain}|}Entropy(R_{Weak}) - \frac{|R_{Strong}|}{|S_{Rain}|}Entropy(R_{Strong})$$

$$0.97 - \frac{3}{5}0 - \frac{2}{5}0$$

$$0.97$$

The diagram below illustrated our decision tree model