target.
↓

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Entropy $H(t) = -\sum_{l \in level(t)} [P(t=l) \times \log(P(t=l))]$

Compute IG. (information gain)

1. compute the entropy of original dataset wrt. target.
2. For each feature, sum entropy of each set. (weighted sum)

~~3.~~ $$rem(d, D) = \sum_{l \in level(d)} \frac{|D_{d=l}|}{|D|} \times H(t, D_{d=l})$$

3. Compute IG: $IG(d, D) = H(t, D) - rem(d, D)$

Stopping criteria:

1. All instance have same target
2. No more features to use. (All features have been used in the branch)
3. The subset is empty.

ID3 example

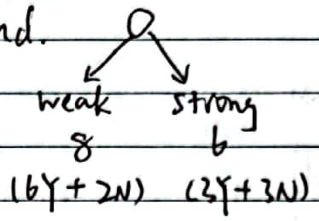
Date 2022/2/18. No.

Target: Decision $\begin{cases} Y: 9 \\ N: 5 \end{cases}$

$$H(t) = -\left(\frac{9}{14} \times \log \frac{9}{14} + \frac{5}{14} \times \log \frac{5}{14}\right) = 0.94$$

Then we need to decide feature to use to split the node.

1. Wind.



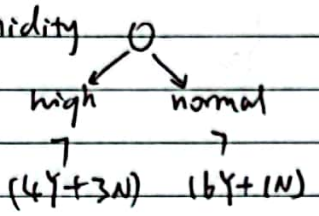
weak subset. $H = -(\frac{6}{8} \times \log \frac{6}{8} + \frac{2}{8} \times \log \frac{2}{8}) = 0.81$

strong subset: $H = -(\frac{3}{6} \times \log \frac{3}{6} + \frac{3}{6} \times \log \frac{3}{6}) = 1$

$\text{rem}(d,D) = \frac{8}{14} \times 0.81 + \frac{6}{14} \times 1 = 0.89$

$IG = 094 - 0.89 = 0.05$

2. Humidity



high subset $H = -(\frac{4}{7} \times \log \frac{4}{7} + \frac{3}{7} \times \log \frac{3}{7}) = 0.985$

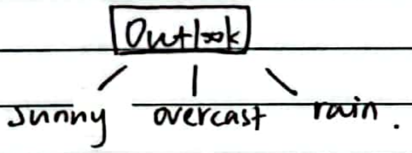
normal subset. $H = -(\frac{6}{7} \times \log \frac{6}{7} + \frac{1}{7} \log \frac{1}{7}) = 0.592$

$\text{rem}(d,D) = \frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.592 = 0.7885$

$IG = .94 - 0.7885 = 0.1515$

3. Temp. IG = 0.029

4. Outlook. IG = 0.246 → largest IG.

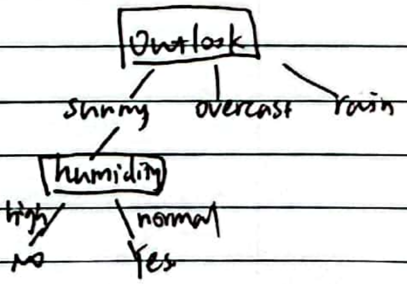


feature set.
temp. humidity. wind.

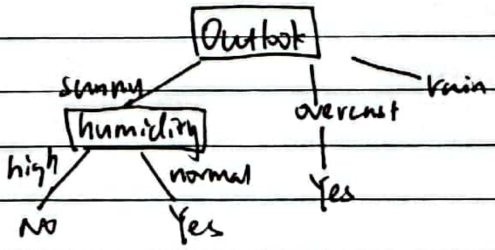
Loop through each subset to further expand.

Outlook = Sunny: 3N+2Y. $H(t) = -(\frac{3}{5} \log \frac{3}{5} + \frac{2}{5} \log \frac{2}{5}) = 0.971$

1. temp. IG = 0.57
2. humidity IG = 0.97 → largest IG
3. wind: IG = 0.017

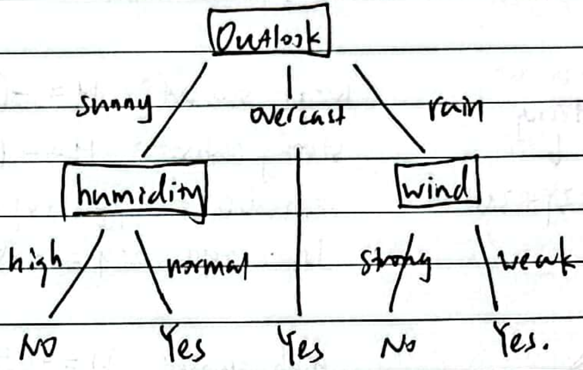


Outlook = Overcast: 4Y. No need to split as all decision is Yes.



Outlook = rain. 3Y+2N $H(t) = -(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}) = 0.971$.

1. temp IG= 0.02
2. humidity IG= 0.02
3. wind IG= 0.97 → largest IG.


Outlook
sunny
overcast
rain
humidity
wind
high
normal
strong
weak
No
Yes
Yes
No
Yes.