Kevin (Chanh P.) Ngo
307385434

Homework 2

1,

* Before Splitting: 14 samples → 7 sunny, 7 rainy
  - Probability of sunny: 7/14 = 1/2
  - Probability of rainy: 7/14 = 1/2

→ Entropy before Splitting:

$$H(X) = -\sum_{k=1}^{K} p(X=x_k)\log_2 p(X=x_k)$$

$$= -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1 \text{ bit}$$

* Split based on Wind (2 branches: Yes and No)
  - Windy (Yes): 6 samples → 1 sunny, 5 rainy
    - Probability of sunny: 1/6
    - Probability of rainy: 5/6
    - → Entropy of windy: $H(X) = -\left(\frac{1}{6}\log_2\left(\frac{1}{6}\right) + \frac{5}{6}\log_2\left(\frac{5}{6}\right)\right) = 0.65$ bit
  - Not Windy (No): 8 samples → 6 sunny, 2 rainy
    - Probability of sunny: 6/8 = 3/4
    - Probability of rainy: 2/8 = 1/4
    - → Entropy of Not Windy: $H(X) = -\left(\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 0.81$ bit

→ Weighted Average: $E(H(x)) = \left(\frac{6}{14} \times 0.65 + \frac{8}{14} \times 0.81\right) = 0.74$ bit

* Split based on Humidity (3 branches: high, mild, low)
  - high: 5 samples → 1 sunny, 4 rainy ⇒ P(Sunny) = 1/5; P(rainy) = 4/5
    - → Entropy of high: $H(X) = -\left(\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \left(\frac{4}{5}\right)\log_2\left(\frac{4}{5}\right)\right) = 0.72$ bit
  - mild: 5 samples → 3 sunny, 2 rainy ⇒ P(sunny) = 3/5; P(rainy) = 2/5
    - → Entropy of mild: $H(X) = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) = 0.97$ bit

→ Weighted Average: $E(H(x)) = ($

⑧ Split based on Humidity → 3 branches: high, mild, low

- high: 5 samples → 1 sunny, 4 rainy → $p(\text{sunny}) = 1/5$; $p(\text{rainy}) = 4/5$

→ Entropy of high: $H(X) = -\left(\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right) = 0.72$ bit

- mild: 5 samples → 3 sunny, 2 rainy → $p(\text{sunny}) = 3/5$; $p(\text{rainy}) = 2/5$

→ Entropy of mild: $H(X) = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) = 0.97$ bit

- low: 4 samples → 3 sunny, 1 rainy → $p(\text{sunny}) = 3/4$; $p(\text{rainy}) = 1/4$

→ Entropy of low → $H(X) = -\left(\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 0.81$ bit.

→ Weighted Average: $E(H(x)) = \left(\frac{5}{14} \times 0.72 + \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0.81\right) = 0.84$ bit.

⑧ Split based on Temp → 3 branches: high, mild, low

- high: 5 samples → 5 sunny, 0 rainy → $p(\text{sunny}) = 5/5 = 1$; $p(\text{rainy}) = 0/5 = 0$

→ Entropy of high: $H(X) = -\left(1 \times \log_2(1) + 0 \times \log_2(0)\right) = 0$ bit

- mild: 4 samples → 2 sunny, 2 rainy → $p(\text{sunny}) = 2/4 = 1/2$; $p(\text{rainy}) = 2/4 = 1/2$

→ Entropy of mild: $H(X) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$ bit

- low: 5 samples → 0 sunny, 5 rainy → $p(\text{sunny}) = 0$; $p(\text{rainy}) = 1$.

→ Entropy of low: $H(X) = -\left(0 \times \log_2(0) + 1 \times \log_2(1)\right) = 0$ bit

→ Weighted Average: $E(H(X)) = \left(\frac{5}{14} \times 0 + \frac{4}{14} \times 1 + \frac{5}{14} \times 0\right) = 0.29$ bit

⑧ Comparisons between features:

	Temp	Humidity	Wind
Entropy before split:	1	1	1
Entropy after split:	0.29	0.84	0.74
Information Grain	1 − 0.29 = 0.71	0.16	0.26

⇒ The best feature to put on the top of the tree is Temp (it gains the most information based on the above table)