Decision Tree.

Entropy of a loaded coin with probability p of heads:

$$-P\log_2 P - (1-P)\log_2(1-P)$$

Case 1. $P(\text{tail}) = 8/9$, $P(\text{head}) = 1/9$

$\Rightarrow -1/9\log_2 1/9 - (1-1/9)\log_2(1-1/9) = -1/9\log_2 1/9 - 8/9\log_2 8/9$

$\Rightarrow -8/9\log_2 8/9 - (1-8/9)\log_2(1-8/9) = -8/9\log_2 8/9 - 1/9\log_2 1/9$ Same

$\Rightarrow 0.5$

Case 2 $P(\text{tail}) = 4/9$, $P(\text{head}) = 5/9$

$\Rightarrow -5/9\log_2 5/9 - 4/9\log_2 4/9 = 0.99.$

Remark: based on Case 1 & 2, the more "impure", "uncertain", "chaotic", the bigger the entropy

For more than binary outcomes:

$$H(Y) = -\sum_{y\in Y} P(y)\log_2 P(y)$$

Joint entropy

$$H(X,Y) = -\sum_x \sum_y P(x,y)\log_2 P(x,y).$$

X \ Y	cloudy	Not cloudy
Rain	24/100	1/100
Not Raining	25/100	50/100

Observe: the probabilities add up to 1.

$$\begin{aligned}
H(X,Y) &= -\sum_x P(x, \text{cloudy})\log_2 P(x, \text{cloudy}) + P(x, \text{not cloudy})\log_2 P(x, \text{not cloudy}) \\
&= -P(\text{rain}, \text{cloudy})\log_2 P(\text{rain}, \text{cloudy}) - P(\text{rain}, \neg\text{cloudy})\log_2 P(\text{rain}, \neg\text{cloudy}) \\
&\quad - P(\neg\text{rain}, \text{cloudy})\log_2 P(\neg\text{rain}, \text{cloudy}) \\
&\quad - P(\neg\text{rain}, \neg\text{cloudy})\log_2 P(\neg\text{rain}, \neg\text{cloudy}) \\
&= -24/100\log_2 24/100 - 1/100\log_2 1/100 - 25/100\log_2 25/100 - 50/100\log_2 50/100 \\
&= 1.56 \text{ bits}
\end{aligned}$$

Specific Condition Entropy $H(Y|X=x) = -\sum_y P(y|x) \log_2 P(y|x)$

$$P(y|x) = \frac{P(x,y)}{P(x)} \qquad P(x) = \sum_y P(x,y)$$

$$\Rightarrow P(rain) = P(rain, cloudy) + P(\neg rain, \neg cloudy)$$

$$H(Y|X=rain) = -\sum_y P(y|rain) \log_2 P(y|rain)$$

$$= -P(cloudy|rain) \log_2 P(cloudy|rain)$$
$$- P(\neg cloudy|rain) \log_2 P(\neg cloudy|rain)$$

$$P(cloudy|rain) = \frac{P(cloudy, rain)}{P(rain)} = \frac{24}{100} \Big/ \frac{24+1}{100} = \frac{24}{25}$$

$$P(\neg cloudy|rain) = \frac{P(\neg cloudy, rain)}{P(rain)} = \frac{1}{100} \Big/ \frac{24+1}{100} = \frac{1}{25}$$

$$H(Y|X=rain) = -24/25 \log_2 24/25 - 1/25 \log_2 1/25 \approx 0.24 \text{ bits}$$

Conditional Entropy

$$H(Y|X) = \sum_x P(x) H(Y|X=x) = -\sum_x \sum_y P(x,y) \log_2 P(x)$$

$$\Rightarrow -\sum_x \sum_y \underbrace{P(x) P(y|x)}_{P(x,y)} \log_2 P(y|x)$$

What is the entropy of cloudiness Y, given the variable X?

$$H(Y|X) = \sum_x P(x) H(Y|X=x) = P(rain) H(Y|X=rain) + P(\neg rain) H(Y|\neg rain)$$

$$= \left(\frac{24}{100} + \frac{1}{100}\right) H(Y|X=rain) + \left(\frac{25}{100} + \frac{50}{100}\right) H(Y|\neg rain)$$

$$= \frac{1}{4} H(Y|rain) + \frac{3}{4} H(Y|\neg rain) \approx 0.75 \text{ bits}$$

Information Gain $IG(Y|X) = H(Y) - H(Y|X)$

$\Rightarrow$ the information we gained on the observation of y after observing how Y "fare" when we have knowledge about X.

Picture at slides 29.

$H(Y) = -\frac{2}{7}\log_2 \frac{2}{7} - \frac{5}{7}\log_2 \frac{5}{7}$

$H(Y|left) = 0.81, \quad H(Y|right) = 0.92$

$IG(split) = 0.86 - (\frac{4}{7} \cdot 0.81 + \frac{3}{7} \cdot 0.92) = 0.006$

Picture at slides 30

$H(Y|left) = 0, \quad H(Y|right) = 0.97$

$IG(split) = 0.86 - (\frac{2}{7} \cdot 0 + \frac{5}{7} \cdot 0.97) = 0.71$

$\Rightarrow$ the split at 30 is better than the split at 29 because a better information gain.