Entropy and Information Gain Notes

Entropy:

- Entropy provides a measurement of uncertainity associated with a random variable or random process.

  Note: This is the definition of entropy under / in the context of information theory.

- For a discrete r.v. X with possible outcomes $x_1, x_2, \ldots, x_n$ which occur with probability $P(x_1), P(x_2), \ldots, P(x_n)$, the entropy of X, denoted as $H(x)$, is defined as:

$$H(x) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

  Note: $H(x) = E\left[-\log_2 P(x)\right]$, where E is expected value.

- Note: When $P(x_i) = 0$, for some x_i, we take $P(x_i) \log_2 P(x_i)$ to be 0, which is consistent with its limit.
  I.e. $\lim_{p \to 0} p \log(p) = 0$

- E.g. Suppose we flip a fair coin.

$$p_1 = p_2 = \tfrac{1}{2}$$

$$\begin{aligned} H(x) &= -(\cancel{2})\left(\tfrac{1}{\cancel{2}} \log_2\left(\tfrac{1}{2}\right)\right) \\ &= -\log_2\left(\tfrac{1}{2}\right) \\ &= -\left(\log_2(1) - \log_2(2)\right) \\ &= -(-1) \\ &= 1 \end{aligned}$$

- E.g. Say we toss an unfair coin now. Suppose the probability of getting heads is 70%. Then, the entropy becomes:

$$H(x) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

$$= -(0.7 \cdot \log_2 (0.7) + 0.3 \cdot \log_2 0.3)$$

$$\approx 0.8816$$ ← Since one side comes up more frequently, there is reduced uncertainty and hence entropy.

- E.g. The entropy of rolling a fair die is:

$$H(x) = -\sum_{i=1}^{n} P(x_i) \cdot \log_2 P(x_i)$$

$$= -\cancel{(6)}\cancel{(\tfrac{1}{6})}(\log_2 (\tfrac{1}{6}))$$

$$= -(\log_2 (1) - \log_2 (6))$$

$$= \log_2 (6)$$

$$= 2.58$$ ← Since the probability of rolling a die (1/6) is smaller than the prob of flipping a coin (1/2), its entropy will be higher.

- We also have conditional entropy.
- $H(x|y) = -\sum_{i,j} P(x_i, y_j) \log_2 P(x_i | y_j)$

$$= -\sum_{j} P(y_j) \sum_{i} P(x_i|y_j) \log_2 P(x_i|y_j)$$

$$= \sum_{j} P(y_j) H(x|y_j)$$

Mutual Information:

- Mutual information is a measure of the info shared by 2 r.v.'s.
  I.e. It is a measure of how much about the state of one such var is known when it is conditioned on the state of the other.

- $I(x;y) = H(x) - H(x|y)$
  $\qquad\quad\, = H(y) - H(y|x)$

- Also called information gain.