CS3236 Introduction to Information Theory

AY2022/23 Semester 2 · Prepared by Tian Xiao *@snoidetx*

1 Information Measures

Information of an Event: If event A occurs with probability p, then we have Information$(A) = \psi(p) = \log_b \frac{1}{p}$.
- When $b = 2$, information is measured in bits.
- Axiomatic view of information:
  - ▶ Non-negativity: $\psi(p) \geq 0$;
  - ▶ Zero for definite events: $\psi(1) = 0$;
  - ▶ Monotonicity: $p \leq p' \Rightarrow \psi(p) \geq \psi(p')$;
  - ▶ Continuity: $\psi(p)$ is continuous in p;
  - ▶ Additivity under independence: $\psi(p_1 p_2) = \psi(p_1) + \psi(p_2)$.

Information of a Random Variable - Entropy: Let X be a discrete random variable with probability mass function P_X. The *Shannon entropy* is the average of information we learn from observing $X = x$:

$$H(X) = \mathbb{E}_{X \sim P(X)}\left[\psi(X = x)\right] = \sum_x P_X(x) \log_2 \frac{1}{P_X(x)}.$$

- $0 \log \frac{1}{0} = 0$.
- Entropy measures the information of uncertainty in X.
- Examples:
  - ▶ Binary source: $H(x) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$.
  - ▶ Uniform source: $H(x) = \log |\mathcal{X}|$.
- Axiomatic view of entropy: Suppose that X is a discrete random variable taking N values with probabilities $\mathbf{p} = \{p_1, \cdots, p_N\}$. Consider an information measure of the form $\Psi(\mathbf{p}) = \Psi(p_1, \cdots, p_N)$:
  - ▶ Continuity: $\Psi(\mathbf{p})$ is continuous as a function of $\mathbf{p}$;
  - ▶ Uniform case: If $\forall i\ \left[p_i = \frac{1}{N}\right]$, then $\Psi(\mathbf{p})$ is increasing in N;
  - ▶ Successive decisions:

$$\Psi(p_1, \cdots, p_N) = \Psi(p_1 + p_2, p_3, \cdots, p_N) + (p_1 + p_2)\Psi\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

- Properties of entropy:
  - ▶ Non-negativity: $H(X) \geq 0$;
  - ▶ Uppon bound: $H(X) \leq \log_2 |X|$;
  - ▶ Chain rule (2 var): $H(X, Y) = H(X) + H(Y|X)$;
  - ▶ Chain rule (n var): $H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \cdots, X_{i-1})$;
  - ▶ Conditioning reduces entropy: $H(X|Y) \leq H(X)$ with equality if and only if X and Y are independent;
  - ▶ Sub-additivity: $H(X_1, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_n)$.
- Variations:
  - ▶ Joint entropy:

$$\begin{aligned} H(X, Y) &= \mathbb{E}_{(X,Y) \sim P(X,Y)}\left[\psi(X = x, Y = y)\right] \\ &= \sum_{x,y} P_{XY}(x, y) \log_2 \frac{1}{P_{XY}(x, y)}. \end{aligned}$$

  - ▶ Conditional entropy:

$$\begin{aligned} H(Y|X) &= \mathbb{E}_{(X,Y) \sim P(X,Y)}\left[\psi(Y = y | X = x)\right] \\ &= \sum_{x,y} P_{XY}(x, y) \log_2 \frac{1}{P_{Y|X}(y|x)} \\ &= \sum_x P_X(x) H(Y | X = x). \end{aligned}$$

KL Divergence: For two PMFs P and Q on a finite alphabet $\mathcal{X}$, the *Kullback-Leibler (KL) divergence* (also known as *relative entropy*) is given by

$$D(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} = \mathbb{E}_{X \sim P}\left[\log_2 \frac{P(x)}{Q(x)}\right].$$

- $D(P||Q) \geq 0$ with equality if and only if $P = Q$.

Information between Random Variables - Mutual Information:

$$I(X; Y) = H(Y) - H(Y|X).$$

- Terminologies:
  - ▶ $H(Y)$: *a priori* uncertainty in Y;
  - ▶ $H(Y|X)$: Remaining uncertainty in Y after observing X;
  - ▶ $I(X; Y)$: Amount of information we learn about Y after observing X.
- Properties of mutual information:
  - ▶ Alternative forms:

$$\begin{aligned} I(X; Y) &= D(P_{XY} || P(X) \times P(Y)) \\ &= \sum_{x,y} P_{XY}(x, y) \log_2 \frac{P_{XY}(x, y)}{P_X(x) P_Y(y)} \\ &= \sum_{x,y} P_{XY}(x, y) \log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)} \end{aligned}$$

  - ▶ Symmetry: $I(X; Y) = I(Y; X) = H(X) + H(Y) - H(X, Y)$.
  - ▶ Non-negativity: $I(X; Y) \geq 0$ with equality if and only if X and Y are independent;
  - ▶ Upper bounds: $I(X; Y) \leq H(X)$; $I(X; Y) \leq H(Y)$;
  - ▶ Chain rule: $I(X_1, \cdots, X_n | Y) = \sum_{i=1}^{n} I(X_i; Y | X_1, \cdots, X_{i-1})$;
  - ▶ Data processing inequality: If $X \perp Z | Y$, $I(X; Z) \leq I(X; Y)$.
  - ▶ Partial sub-additivity: If $Y_1, \cdots, Y_n$ are conditionally independent given $X_1, \cdots, X_n$, and Y_i depends on $X_1, \cdots, X_n$ only through X_i, then $I(X_1, \cdots, X_n; Y_1, \cdots, Y_n) \leq \sum_{i=1}^{n} I(X_i; Y_i)$.
- Variations:
  - ▶ Joint version:

$$I(X_1, X_2; Y_1, Y_2) = H(Y_1.Y_2) - H(Y_1, Y_2 | X_1, X_2).$$

  - ▶ Conditional version:

$$I(X; Y|Z) = H(Y|Z) - H(Y|X, Z).$$

2 Symbol-wise Source Coding

Symbol-wise Coding: Symbol-wise source coding maps each $x \in \mathcal{X}$ to some binary sequence $C(x)$. The length of this sequence is denoted by $\ell(x)$. The average length of a code $C(\cdot)$ is given by

$$L(C) = \sum_{x \in \mathcal{X}} P_X(x) \ell(x).$$

- Non-singular: $x \neq x' \Rightarrow C(x) \neq C(x')$.
- Uniquely decodable: A code $C(\cdot)$ is said to be *uniquely decodable* if no two **sequences** of symbols in $\mathcal{X}$ are coded to the same concatenated binary sequence.
- Prefix-free: A code $C(\cdot)$ is said to be *prefix-free* if no codeword is a prefix of any other.

Kraft's Inequality: Any prefix-free code $C(\cdot)$ that maps each $x \in \mathcal{X}$ to a codeword of length $\ell(x)$ must satisfy

$$\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1.$$

- Existence Property: If a given set of integers $\{\ell(x)\}_{x \in \mathcal{X}}$ satisfies $\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1$, then it is possible to construct a prefix-free code that maps each $x \in \mathcal{X}$ to a codeword of length $\ell(x)$.

Entropy Bound: For $X \sim P_X$ and any prefix-free code $C(\cdot)$, the expected length satisfies

$$L(C) \geq H(X),$$

with equality if and only if $P_X(x) = 2^{-\ell(x)}$ for all $x \in \mathcal{X}$.

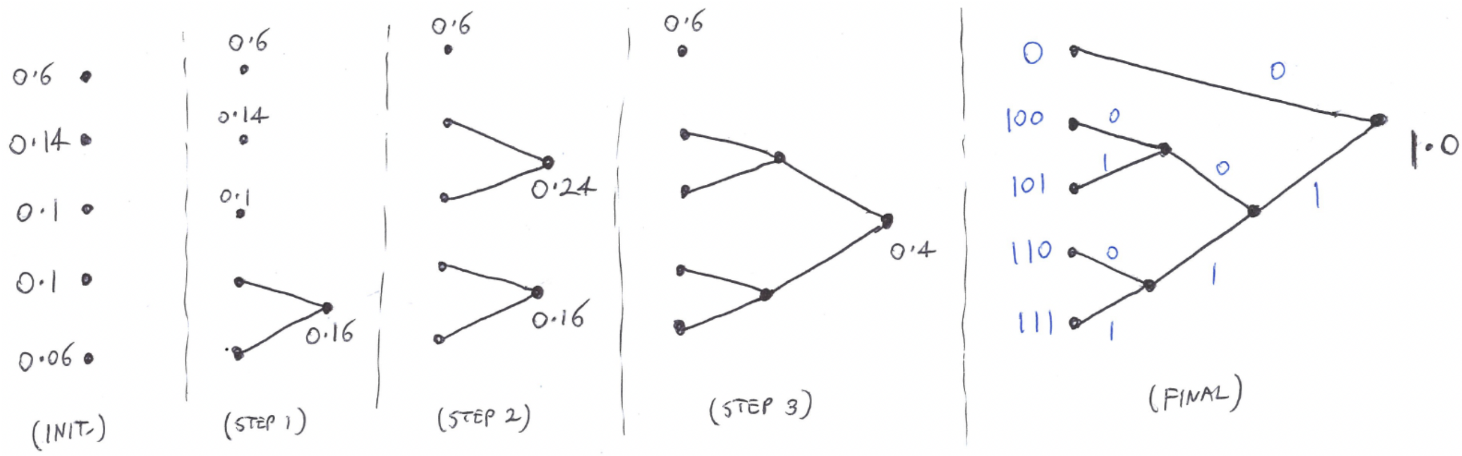
Shannon-Fano Code: $\ell(x) = \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil$.
- $H(X) \leq L(C) < H(X) + 1$.
- If the true distribution is P_X but the lengths are chosen according to Q_X, then the Shannon-Fano code satisfies

$$H(X) + D(P_X || Q_X) \leq L(C) \leq H(X) + D(P_X || Q_X) + 1.$$

Huffman Code: Construct a tree as follows:

1. List the symbols of $\mathcal{X}$ from highest probability to lowest.
2. Draw a branch connecting the two symbols with the lowest probability, and label the merged point with the sum of the two associated probabilities.
3. Repeat the first two steps until everything has merged to a single point with total probability 1.

- No uniquely decodable symbol code can achieve a smaller average length $L(C)$ than the Huffman code.
- $H(X) \leq L(C) < H(X) + 1$.


0·6
0·14
0·1
0·1
0·06
(INIT.)
0·6
0·14
0·1
0·16
(STEP 1)
0·6
0·24
0·16
(STEP 2)
0·6
0·4
(STEP 3)
0
100
101
110
111
0
0
1
0
1
0
1
1
1·0
(FINAL)