CS258 Information Theory Homework 3

Zhou Litao 518030910407 F1803016

March 17, 2020

Exercise 1 Prove that under the constraint that $X \to Y \to Z$ forms a Markov Chain, $X \perp Y | Z$ and $X \perp Z$ imply $X \perp Y$.

Proof. From $X \perp Y|Z$, we have $I(X;Y|Z) = 0$. From $X \perp Z$, we have $I(X;Z) = 0$. It follows that

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) && \text{(Unfold by definition of mutual information)} \\ &= H(X) - H(X|Z) + H(X|Z) - H(X|Y) \\ &= H(X) - H(X|Z) + H(X|Z) - H(X|Y,Z) && \text{(Markov Chain: } p(x|y) = p(x|y,z)) \\ &= I(X;Z) + I(X;Y|Z) = 0 && \text{(Fold by definition of mutual information)} \end{aligned} \tag{1}$$

,which implies that $X \perp Y$. □

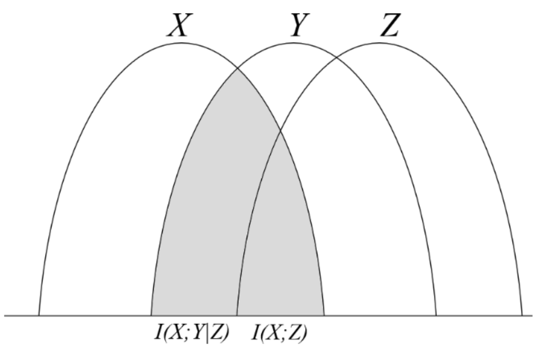

Figure 1: Venn Diagram of Exercise 1

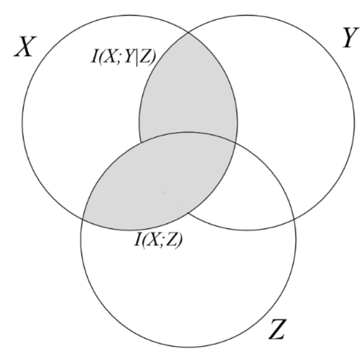

Figure 2: Venn Diagram of Exercise 2

Exercise 2 Prove that the implication in Exercise 1 continues to be valid without the Markov Chain constraint

Proof.

$$\begin{aligned} I(X;Y) &= I(X;Y|Z) + (I(X;Y) - I(X;Y|Z)) && \text{(Note } X \perp Y|Z \to I(X;Y|Z) = 0) \\ &= H(X) - H(X|Y) - (H(X|Z) - H(X|Y,Z)) && \text{(Fold by definition of mutual information)} \\ &= (H(X) - H(X|Z)) - (H(X|Y) - H(X|Y,Z)) && \text{(Unfold by definition of mutual information)} \\ &= I(X;Z) - I(X;Z|Y) && \text{(Note } X \perp Y \to I(X;Y) = 0) \\ &= -I(X;Z|Y) \le 0 && \text{(Nonnegative conditional mutual information)} \end{aligned} \tag{2}$$

On the other hand, $I(X;Y) \ge 0$. Hence $I(X;Y)$ must be zero. That is to say, $X \perp Y$. □

Exercise 3 Prove that $Y \perp Z|T$ implies $Y \perp Z|(X,T)$ conditioning on $X \to Y \to Z \to T$.

Proof.

$$
\begin{aligned}
I(Y;Z|X,T) &= H(Y|X,T) - H(Y|Z,X,T) && \text{(Unfold mutual information)}\\
&= H(X,Y,T) - H(X,T) - H(X,Y,Z,T) + H(X,Z,T) && \text{(Unfold conditional entropy)}\\
&= (H(X,Y,T) - H(X,Y,Z,T)) - (H(X,T) - H(T))\\
&\quad +(H(X,Z,T) - H(Z,T)) - H(T) + H(Z,T)\\
&= -H(Z|X,Y,T) - H(X|T) + H(X|Z,T) + H(Z|T) && \text{(Fold conditional entropy)}\\
&= (H(Z|T) - H(Z|Y,T)) - (H(X|T) - H(X|Z,T)) && \text{(Markov Chain: } p(z|x,y,t) = p(z|y,t))\\
&= I(Y;Z|T) - I(X;Z|T) && \text{(Note } Y \perp Z|T \to I(Y;Z|T) = 0)\\
&= -I(X;Z|T) \le 0
\end{aligned}
\tag{3}
$$

On the other hand, $I(Y;Z|X,T) \ge 0$. Hence $I(Y;Z|X,T)$ must be zero. That is to say, $Y \perp Z|(X,T)$. ☐

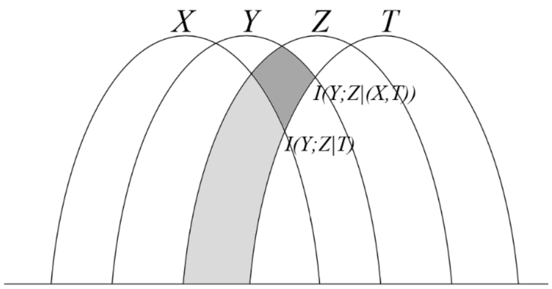

Figure 3: Venn Diagram of Exercise 3

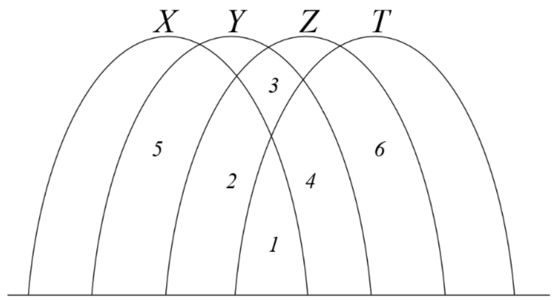

Figure 4: Venn Diagram of Exercise 4

Exercise 4 Let $X \to Y \to Z \to T$ form a Markov Chain. Determine which of the following always hold:

1. $I(X;T) + I(Y;Z) \ge I(X;Z) + I(Y;T)$
2. $I(X;T) + I(Y;Z) \ge I(X;Y) + I(Z;T)$
3. $I(X;Y) + I(Z;T) \ge I(X;Z) + I(Y;T)$

Solution. Inequality (1) and (3) always hold. We illustrate the answer through the venn diagram shown in Figure 4, where area $1 \sim 6$ is respectively represented by $I(X;T), I(X;Z|T), I(Y;Z|(X,T)), I(Y;T|X), I(X;Y|Z), I(Z;T|Y)$.

1. The inequality can be rewritten in form of areas as

$$1 + (1 + 2 + 3 + 4) \ge (1 + 2) + (1 + 4).$$

   Since $I(Y;Z|(X,T)) \ge 0$ has been proved in Exercise 3, the inequality holds.

2. The inequality can be rewritten in form of areas as

$$1 + (1 + 2 + 3 + 4) \ge (1 + 2 + 5) + (1 + 4 + 6).$$

   We can't determine the relation between $I(Y;Z|(X,T))$ (Area 3) and $I(X;Y|Z) + I(Z;T|Y)$ (Area 5 and 6) except that they are both nonnegative. The inequality will not always hold.

3. The inequality can be rewritten in form of areas as

$$(1+2+5)+(1+4+6) \geq (1+2)+(1+4).$$

Since $I(X;Y|Z)+I(Z;T|Y) \geq 0$ can be proved by the nonnegativity of conditional mutual information, the inequality holds. Furthermore, the conclusion can also be derived from the data-processing inequality of Markov Chain with $I(X;Y) \geq I(X;Z)$ and $I(Z;T) \geq I(Y;T)$

□

Exercise 5 (Drawing with and without replacement) An urn contains r red, w white, and b black balls. Which has higher entropy, drawing $k \geq 2$ balls from the urn with replacement or without replacement? Set it up and show why. (There is both a difficult way and a relatively simple way to do this.)

Solution. We use $X_i \in \{\text{red, white, black}\}$ to identify the result of the i-th drawing. No matter with replacement or without replacement, the distributions of a single arbitrary variable X_i are the same.

$$\begin{array}{cccc} X_i & red & white & black \\ p(x) & \dfrac{r}{r+w+b} & \dfrac{w}{r+w+b} & \dfrac{b}{r+w+b} \end{array} \tag{4}$$

With replacement, the previous result won't interfere with the present drawing. Hence we have

$$H\left(X_i|X_{i-1},\ldots,X_1\right) = H\left(X_i\right)$$

. It follows that

$$H(X_1,X_2,...,X_k) = \sum_{i=1}^{k} H\left(X_i|X_{i-1},\ldots,X_1\right) = \sum_{i=1}^{k} H\left(X_i\right) \quad \text{with replacement} \tag{5}$$

Without replacement, we only have

$$H(X_1,X_2,...,X_k) = \sum_{i=1}^{k} H\left(X_i|X_{i-1},\ldots,X_1\right) \quad \text{without replacement} \tag{6}$$

Note that in Equation 5 and Equation 6, all the single-variable entropies are of the same value. By condition-reduce-entropy theorem we know that

$$H\left(X_i|X_{i-1},\ldots,X_1\right) \leq H(X_i) \quad \text{for any } i$$

Since the equality holds if and only if X_i are mutually independent, which is not true in this problem, it follows that the entropy will be larger with replacement. □

Exercise 6 (Metric) A function $\rho(x,y)$ is a metric if for all x,y,

- $\rho(x,y) \geq 0$.
- $\rho(x,y) = \rho(y,x)$.
- $\rho(x,y) = 0$ if and only if $x = y$.
- $\rho(x,y) + \rho(y,z) \geq \rho(x,z)$.

1. Show that $\rho(X,Y) = H(X|Y) + H(Y|X)$ satisfies the first, second, and fourth properties above. If we say that $X = Y$ if there is a one-to-one function mapping from X to Y, the third property is also satisfied, and $\rho(X,Y)$ is a metric.

2. Verify that $\rho(X,Y)$ can also be expressed as

$$\begin{aligned} \rho(X,Y) &= H(X) + H(Y) - 2I(X;Y) \\ &= H(X,Y) - I(X;Y) \\ &= 2H(X,Y) - H(X) - H(Y) \end{aligned} \tag{7}$$

Proof.

1. - Note that $H(X|Y) \geqslant 0, \quad H(Y(X) \geqslant 0 \Rightarrow \quad \rho(X,Y) \geqslant 0$
   - By unfolding the definition it's easy to see $H(X|Y) \neq H(Y|X) = H(Y|X) + H(X|Y)$
   - If $X = Y$, there exists a one-to-one mapping, i.e. $\rho(X,Y) = H(X|Y) + H(Y|X) = 0$
     On the other hand, if $\rho(X,Y) = H(X|Y) + H(Y|X) = 0$ Since $H(X|Y) \geqslant 0, \quad H(Y(X) \geqslant 0$ By the conclusion in Exercise 2, Assignment 1, there exists a one-to-one mapping between X and Y.
   - The conclusion can be derived using condition-reduce-entropy and nonnegativity of conditional entropy.

$$\begin{aligned} H(X|Y) + H(Y|X) + H(Y|Z) + H(Z|Y) &\geqslant H(X|Y,Z) + H(Y|X) + H(Y|Z) + H(Z|Y,X) \\ &= H(X,Y|Z) + H(Z,Y|X) \\ &= H(X|Z) + H(Y|X,Z) + H(Z|X) + H(Y|Z,X) \\ &\geqslant H(X|Z) + H(Z|X) = \rho(X,Z) \end{aligned} \tag{8}$$

2. $$\begin{aligned} \rho(X,Y) &= H(X|Y) + H(Y|X) \\ &= H(X) - I(X;Y) + H(Y) - I(X;Y) \\ &= H(X) + H(Y) - 2I(X;Y) && (\star) \\ &= (H(X) - I(X;Y) + H(Y)) - I(X;Y) \\ &= H(X,Y) - I(X;Y) && (\star) \\ &= H(X,Y) - (H(X) + H(Y) - H(X,Y)) \\ &= 2H(X,Y) - H(X) - H(Y) && (\star) \end{aligned} \tag{9}$$

   The expressions required by the problem have been labeled with $(\star)$ in the derivation.

□

Exercise 7 (Entropy of a disjoint mixture) Let X_1 and X_2 be discrete random variables drawn according to probability mass functions $p_1(\cdot)$ and $p_2(\cdot)$ over the respective alphabets $X_1 = \{1, 2, \cdots, m\}$ and $X_2 = \{m+1, \cdots, n\}$. Let

$$X = \begin{cases} X_1 & \text{with probability } \alpha \\ X_2 & \text{with probability } 1-\alpha \end{cases} \tag{10}$$

1. Find $H(X)$ in terms of $H(X_1), H(X_2)$ and α.
2. Maximize over α to show that $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$ and interpret using the notion that $2^{H(X)}$ is the effective alphabet size.

Solution.

1. We calculate $H(X)$ by unfolding the definition of entropy.

$$\begin{aligned} H(X) &= -\sum_{x \in X_1} \alpha p_1(x) \log \alpha p_1(x) - \sum_{x \in X_2} (1-\alpha) p_2(x) \log(1-\alpha) p_2(x) \\ &= -\alpha \log \alpha \sum_{x \in X_1} p_1(x) - (1-\alpha)\log(1-\alpha) \sum_{x \in X_2} p_2(x) + \alpha H(X_1) + (1-\alpha) H(X_2) \\ &= -\alpha \log \alpha - (1-\alpha)\log(1-\alpha) + \alpha H(X_1) + (1-\alpha) H(X_2) \end{aligned} \tag{11}$$

2. We consider $H(X)$ to be a function over α. Note that $g(\alpha) = -\alpha \log(\alpha)$ is a concave function, and some affine transformation over α and linear components won't interfere with the concavity. The function of $H(x)$ is a concave function.

We can get the maximal value by calculating the derivative of $H(X)$ over α.

$$\begin{aligned}\frac{dH(X)}{d\alpha} &= -\frac{1}{d\alpha}\left(\frac{\alpha\ln\alpha}{\ln 2} + \frac{(1-\alpha)\ln(1-\alpha)}{\ln 2} - \alpha H(X_1) - (1-\alpha)H(X_2)\right) \\ &= -\frac{1+\ln\alpha}{\ln 2} - \frac{-1-\ln(1-\alpha)}{\ln 2} + H(X_1) - H(X_2) := 0\end{aligned} \tag{12}$$

The maximal value is obtained at the derivative to be 0.

$$\begin{aligned}-\ln\alpha + \ln(1-\alpha) &= \ln 2\,(H(X_2) - H(X_1)) \\ \ln\frac{1-\alpha}{\alpha} &= \ln 2\,(H(X_2) - H(X_1)) \\ \frac{1-\alpha}{\alpha} &= 2^{H(X_2)-H(X_1)} \\ \alpha &= \frac{2^{H(X_1)}}{2^{H(X_2)} + 2^{H(X_1)}}\end{aligned} \tag{13}$$

The optimal solution is in the domain, so the maximal value can be obtained. By substituting the α value into $2^{H(X)}$ we can obtain its upper bond.

$$\begin{aligned}2^{H(X)} &= 2^{-\alpha\log\alpha-(1-\alpha)\log(1-\alpha)+\alpha H(X_1)+(1-\alpha)H(X_2)} \\ &= \alpha^{-\alpha}\cdot(1-\alpha)^{\alpha-1}\cdot\left(2^{H(X_1)}\right)^{\alpha}\cdot\left(2^{H(X_2)}\right)^{1-\alpha} \\ &\le \left(\frac{2^{H(X_1)}}{2^{H(X_2)}+2^{H(X_1)}}\right)^{-\alpha}\cdot\left(\frac{2^{H(X_2)}}{2^{H(X_2)}+2^{H(X_1)}}\right)^{\alpha-1}\cdot\left(2^{H(X_1)}\right)^{\alpha}\cdot\left(2^{H(X_2)}\right)^{1-\alpha} \\ &= \left(2^{H(X_1)}+2^{H(X_2)}\right)\cdot 2^{-\alpha H(X_1)}\cdot 2^{-(1-\alpha)H(X_2)}\cdot\left(2^{H(X_1)}\right)^{\alpha}\cdot\left(2^{H(X_2)}\right)^{1-\alpha} \\ &= 2^{H(X_1)}+2^{H(X_2)}\end{aligned} \tag{14}$$

An interpretation of this conclusion is that $2^{H(X)}$ is the effective alphabet size of X, while $2^{H(X_1)} + 2^{H(X_2)}$ is the sum sizes of the effective alphabets X_1, X_2. The alphabets of X_1 and X_2 do not overlap, with independent distribution, and they add up exactly to the alphabet of X.

If our probability of choice between X_1 and X_2 is in proportion to their effective alphabet size, as the third line in Equation 13 shows, the resulting X will have the effective alphabet size equivalent to the sum of X_1 and X_2.

Otherwise, the unbalanced weight of X_1 and X_2 will reduce the actual effective alphabet size in X, since one variable's excessive occurence will reduce the occurence of the other, so that the latter's effective alphabet size will be less than what it really is.

□

Exercise 8 (Entropy of a sum) Let X and Y be random variables that take on values $x_1, x_2, \cdots, x_r$ and $y_1, \cdots, y_s$, respectively. Let $Z = X + Y$.

- Show that $H(Z|X) = H(Y|X)$. Argue that if X, Y are independent, then $H(Y) \le H(Z)$ and $H(X) \le H(Z)$. Thus, the addition of independent random variables adds uncertainty.
- Give an example of (necessarily dependent) random variables in which $H(X) > H(Z)$ and $H(Y) > H(Z)$.
- Under what conditions does $H(Z) = H(X) + H(Y)$?

Solution.

1. $Z = X + Y$ indicates that any of the two variable can determine the third variable. That is to say,

$$H(X|Y,Z) = H(Y|Z,X) = H(Z|X,Y) = 0$$

. By observing $I(Y;Z|X)$ we have

$$\begin{aligned} I(Y;Z|X) &= H(Y|X) - H(Y|X,Z) \\ &= H(Z|X) - H(Z|X,Y) \end{aligned} \tag{15}$$

,which implies that $H(Y|X) = H(Z|X)$.
If X and Y are independent, $H(X,Y) = H(X) + H(Y)$.

$$\begin{aligned} H(X,Y,Z) &= H(Z|X,Y) + H(X,Y) = H(X) + H(Y) \\ &= H(X|Y,Z) + H(Y,Z) = H(Z|X) + H(X) \\ &= H(Y|X,Z) + H(X,Z) = H(Z|Y) + H(Y) \end{aligned} \tag{16}$$

Equation 16 indicates that $H(Z|X) = H(Y)$ and that $H(Z|Y) = H(X)$. By condition-reduce-entropy theorem we have $H(Z) \geq H(Z|X)$ and $H(Z) \geq H(Z|Y)$. It follows that $H(Y) \leq H(Z)$ and $H(X) \leq H(Z)$.

2. An exampling distribution of X and Y can be

Prob		x	
		0	1
y	0	$\frac{1}{2}$	0
	-1	0	$\frac{1}{2}$

The entropy of X and Y are

$$H(X) = H(Y) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1$$

The distribution of $Z = X + Y$ is $\Pr(Z = 0) = 1$, which results in the entropy $H(Z) = 0 < H(X) = H(Y)$.

3. From $Z = X + Y$ we know $H(Z) = H(Z) - H(Z|X,Y) = I(X,Y;Z)$.

$I(X,Y;Z) = H(X,Y) - H(X,Y|Z)$ indicates that $H(Z) \leq H(X,Y)$. The equality holds if and only if $H(X,Y|Z) = 0$.

Furthermore, $H(X,Y) = H(X)+H(Y)-I(X;Y)$, which implies that $H(X,Y) \leq H(X)+H(Y)$. The equality holds if and only if $I(X;Y) = 0$, i.e. X and Y are independent.

The second equality constraint and the propositions that $H(X|Y,Z) = H(Y|Z,X) = 0$ can ensure the first equality constraint. Therefore, under the condition that X and Y are independent will $H(Z) = H(X)+H(Y)$ hold.

$\square$

Exercise 9 (Data processing) Let $X_1 \to X_2 \to X_3 \to \cdots \to X_n$ form a Markov chain in this order; that is, let

$$p(x_1, x_2, \ldots, x_n) = p(x_1)\, p(x_2|x_1) \cdots p(x_n|x_{n-1})$$

Reduce $I(X_1; X_2, \cdots, X_n)$ to its simplest form.

Solution. By the chain rule of mutual information we have

$$I(X_1; X_2, \ldots, X_n) = \sum_{i=2}^{n} I(X_i; X_1 | X_{i-1}, X_{i-2}, \ldots, X_2)$$

Note that for $i > 2$, we have

$$\begin{aligned} I\left(X_i; X_1 | X_{i-1}, X_{i-2}, \ldots, X_2\right) &= H\left(X_i | X_{i-1}, X_{i-2}, \ldots, X_2\right) - H\left(X_i | X_{i-1}, X_{i-2}, \ldots, X_2, X_1\right) \\ &= H(X_i|X_{i-1}) - H(X_i|X_{i-1}) = 0 \quad \text{(Markov Chain)} \end{aligned} \tag{17}$$

It follows that $I\left(X_1; X_2, \ldots, X_n\right) = I\left(X_1; X_2\right)$.

□

Exercise 10 (Infinite entropy) This problem shows that the entropy of a discrete random variable can be infinite. Let $A = \sum_{n=2}^{\infty}(n \log^2 n)^{-1}$. [It is easy to show that A is finite by bounding the infinite sum by the integral of $(x \log^2 x)^{-1}$.] Show that the integer-valued random variable X defined by $\Pr(X = n) = (An \log^2 n)^{-1}$ for $n = 2, 3, \cdots$, has $H(X) = +\infty$.

Proof. By definition of entropy we can calculate that

$$\begin{aligned} H(X) &= -\sum_{n=2}^{\infty} p(n) \log p(n) \\ &= \sum_{n=2}^{\infty} (An \log^2 n)^{-1} \log(An \log^2 n) \\ &= \sum_{n=2}^{\infty} \frac{\log A + \log n + \log^2 n}{An \log^2 n} \\ &= \log A + \sum_{n=2}^{\infty} \frac{1}{An \log n} + \sum_{n=2}^{\infty} \frac{\log^2 n}{An \log^2 n} \end{aligned} \tag{18}$$

As has been indicated by the condition, the first component is finite. The last component will be nonnegative with sufficiently large n. We show that the second component is infinite. Note that

$$0 < \sum_{n=2}^{\infty} \frac{1}{An \log n} < \int_2^{\infty} \frac{\ln 2 dx}{Ax \ln x} = \int_2^{\infty} \frac{\ln 2 d(\ln x)}{A \ln x} = \frac{\ln 2}{A} \ln(\ln x)\Big|_2^{\infty} \to \infty$$

It follows that $H(X) = +\infty$ □