CS3236: Tutorial 1
(Information Measures)

Part I – Entropy (for the coming week(s))

1. [**Example Entropy Calculations**]

   Recall the definition of the binary entropy function,

   $$H_2(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}.$$

   Suppose that $X \sim \text{Bernoulli}(p)$, and

   $$P_{Y|X}(y|x) = \begin{cases} 1-\delta & y = x \\ \delta & y = 1-x. \end{cases}$$

   That is, Y is a "noisy" version of X that is flipped with probability δ.

   (a) Calculate the entropies $H(X)$ and $H(Y)$. Express your answers in terms of the function $H_2(\cdot)$.

   (b) Calculate the conditional entropies $H(Y|X)$ and $H(X|Y)$. Express your answers in terms of the function $H_2(\cdot)$. *(Note: The expression for $H(X|Y)$ is not "neat", and it may help to define the shorthand $q = \mathbb{P}[Y = 1]$ to lighten the notation.)*

2. [**Coin Tossing and Entropy**]

   A fair coin is tossed multiple number of times until we see the first head in the i-th coin toss and then followed by second head in the j-th coin toss.

   Find the probability distribution of the random variable $X = j - i$, and calculate its entropy.

   (Hint: You may wish to use the identities $\sum_{r=1}^{\infty} 2^{-r} = 1$ and $\sum_{r=1}^{\infty} r2^{-r} = 2$)

3. [**Binary Entropy of Average vs. Average of Binary Entropy**]

   Recall the definition of the binary entropy function, $H_2(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$ for $p \in [0,1]$, which is the entropy of a Bernoulli(p) random variable.

   The purpose of this question is to show that for any parameters $p_1, \dots, p_n$ in the range $[0,1]$, the following holds:

   $$\frac{1}{n} \sum_{j=1}^{n} H_2(p_j) \le H_2\left(\frac{1}{n} \sum_{j=1}^{n} p_j \right).$$

   Some of you may recognize this as an application of Jensen's inequality for concave functions, and such an approach can indeed prove the equation.

   Suppose that someone (possibly yourself!) hasn't heard of Jensen's inequality or concavity, and wants to see a different proof. They are, however, familiar with the fact that conditioning reduces entropy:

   $$H(X|Y) \le H(X).$$

Prove the first display equation above via a suitable choice of X and Y. *(Hint: First define $X_1, \dots, X_n$ to be Bernoulli with parameters $p_1, \dots, p_n$, and then consider a randomly-chosen index from $\{1, \dots, n\}$.)*

4. [**Decomposability of Entropy**]

   (a) For a given probability distribution P , where $P = \{p_1, p_2, ..., p_n\}$, Prove the following equation:

   $$H(P) = H(p_1, 1 - p_1) + (1 - p_1) H\left(\frac{p_2}{1 - p_1}, \frac{p_3}{1 - p_1}, ..., \frac{p_n}{1 - p_1}\right)$$

   showing that the entropy is decomposable. *(Note: Overloading notation slightly, here $H(P)$ and $H(p_1, \dots, p_n)$ denote the usual entropy $H(X)$ for X distributed according to P)*

   (Hint: Let $X \sim P$ and consider the chain rule $H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$ with a carefully-chosen Y)

   (b) A coin whose probability of obtaining heads 2/3 and tails 1/3 is flipped until the first head is obtained. Using the decomposability of the entropy above, what is the entropy of the random variable $X \in \{1, 2, 3, \dots\}$, the number of flips?

5. [**Entropy of a Function**]

   Let X be a random variable taking on a finite number of real values $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, and let Y be a deterministic function of X.

   (a) Prove the inequality $H(X) \geq H(Y)$. *(Note: The optional section of the lecture notes has a proof. Try to give an alternative proof using the fact that different x values mapping to a common y value can only lead to the combined probability being higher; lower probabilities are never produced.)*

   (b) Give an example where $H(X) > H(Y)$, and one where $H(X) = H(Y)$.

   (c) Prove that $H(X_1 + X_2) \leq H(X_1, X_2)$ (here we assume that these random variables are defined on the integers or the reals, so the notion of addition makes sense)

6. [**Bizarre Balance**]

   You are given 10 balls, all of which are equal in weight w except for one which weights $1.01w$. You are also given a bizarre two-pan balance that can report only three outcomes: 'Twice of the left side weight is greater than right side weight' or 'Twice of left side weight is less than right side weight' or 'Twice of left side weight is equal to the right side weight'.

   (a) Argue that at least 3 weighings are needed to guarantee that the odd ball can be identified.

   (b) Design a strategy to determine which is the odd ball while always using 3 weighings or fewer. Note that the choice of the next weighing is allowed to depend on all of the outcomes observed so far.

   (Hint: Try to maximize information of outcomes in each weighing.)

7. [**Alternative Proof of $D_{KL} \geq 0$**]

   Use Jensen's inequality to prove that the relative entropy $D(P||Q)$ satisfies $D(P||Q) \geq 0$ (Gibbs' inequality) with equality only if $P = Q$.

   (Note: Jensen's inequality states that if $\mathbf{X}$ is a random variable (or random vector) and f is a convex function, then $f(\mathbb{E}[\mathbf{X}]) \leq \mathbb{E}[f(\mathbf{X})]$. Moreover, if the function is strictly convex, then equality holds if and only if $\mathbf{X}$ is deterministic (takes some value with probability one). Note that the function $f(u) = -\log(u)$ is strictly convex, since the log function is strictly concave.)

Hints for Part I

1. Mostly substitution into the definition of (conditional) entropy. Recall $H_2(p)$ is the entropy of a $\{0,1\}$-valued (Bernoulli) RV with probability p of being q. Compute $\mathbb{P}[Y=1]$ directly.

2. Look up the geometric distribution if it helps. Otherwise, a direct calculation along with the hint.

3. Let the index J be uniformly random on $\{1,\dots,n\}$, and let X be the corresponding random variable X_J. Use $H(X|J) \le H(X)$ and evaluate both sides.

4. In (a), choose $Y = \mathbf{1}\{X=1\}$, note that $H(X|Y=1)=0$, and apply the definition of conditional probability to $\mathbb{P}[X=x|Y=0]$. In (b), make use of part a, and note that if the first toss is a tail then the resulting distribution of "how many flips left" is back exactly where it started.

5. In (a), first show that $\sum_{x\,:\,y=f(x)} P_X(x)\log_2 P_X(x) \le P_Y(y)\log_2 P_Y(y)$, and then try to evaluate $H(X)$ by writing $\sum_x$ as a double-sum $(\sum\sum(\cdot))$. In (b) just play around with a few examples on a ternary alphabet (i.e., $|\mathcal{X}|=3$). Explain why (c) is a special case of part a.

6. In (a) use a simple counting argument (entropy is not needed). In (b) maximizing $H(Y)$ or minimizing $H(X|Y)$ is a reasonable strategy (in fact, as shown in the solutions, the two are equivalent).

7. Consider the function $f(u) = -\log(u)$, and write $D(P\|Q)$ as an average of $f(\cdot)$.

Part II - Mutual Information (for a later week)

8. [**Three Cards**]

   (a) One card is white on both faces; one is black on both faces; and one is white on one side and black on the other. The three cards are shuffled and their orientations randomized. One card is drawn and placed on the table. The upper face is black. What is the probability that the color of its lower face is white ?
   (b) Does seeing the top face convey information about the color of the bottom face? Discuss the information contents and entropies in this situation. Let the value of the upper face's color be U and the value of the lower face's color be L. Imagine that we draw a random card and learn both U and L. What is the entropy $H(U)$? What is the entropy $H(L)$? What is the mutual information between U and $L, I(U;L)$?

9. [**Chain Rule for Mutual Information**]

   Recall the following definitions and properties.

   Conditional Mutual Information: The conditional mutual information between random variables X and Y given Z is given by $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$. This is a generalization of the equation $I(X;Y) = H(X) - H(X|Y)$.

   Chain rule for Entropy: $H(X,Y) = H(X) + H(Y|X)$.

   Using the chain rule for entropy, prove that $I(X,Y;Z) = I(X;Z) + I(Y;Z|X)$. In addition, show that it implies $I(X,Y;Z) \ge I(X;Z)$.

10. [**General Chain Rule for Mutual Information**]

    Show that $I(X_1,X_2,\dots,X_n\,;Y) = I(X_1\,;Y) + I(X_2\,;Y|X_1) + \cdots + I(X_n\,;Y|X_1,X_2,\dots,X_{n-1}) = \sum_{i=1}^{n} I(X_i\,;Y|X_1,X_2,\dots,X_{i-1})\,.$

11. [**Data Processing Inequality**]

    Recall that the data processing inequality states that if $X \to Y \to Z$ (i.e., X and Z are conditionally independent given Y), then $I(X;Z) \le I(X;Y)$.

(a) Give a proof different from the one in the lecture by applying two different versions of the chain rule to $I(X;Y,Z)$.

(b) When does the data processing inequality hold with equality?

(c) Prove also that if $X \to Y \to Z$ then $I(X;Z) \leq I(Y;Z)$.

12. [**Entropy Distance**]

Let the probability distribution of the joint random variable XYZ be P_{XYZ} and the marginal distributions of RV's XY, YZ, ZX, X, Y and Z be P_{XY}, P_{YZ}, P_{ZX}, P_X, P_Y, P_Z over the alphabets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$.

Let use define the 'entropy distance' between random variables X and Y (given the joint probability distribution of XYZ) to be the difference between their joint entropy and their mutual information:

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

Similarly,

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

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

Prove that the entropy distance satisfies the following: (i) $D_H(X,Y) \geq 0$, (ii) $D_H(X,Y) = D_H(Y;X)$, and (iii) $D_H(X,Z) \leq D_H(X,Y) + D_H(Y,Z)$.

Note that

$$\begin{aligned} D_H(X,Y) &\equiv H(X,Y) \; - \; I(X;Y) \\ &= (H(Y) + H(X|Y)) \; - \; (H(Y) - H(Y|X)) \\ &= H(X|Y) \; + \; H(Y|X). \end{aligned}$$

(Note: The third property is by far the most challenging – the steps used include "conditioning reduces entropy", the chain rule in two different forms, and the non-negativity of (conditional) entropy.)

13. [**Independence and Mutual Information**]

Show that if $X_1, \ldots, X_n$ are mutually independent, then

$$I(X_1, \ldots, X_n; Y_1, \ldots, Y_n) \geq \sum_{i=1}^{n} I(X_i; Y_i).$$

In addition, show that if given X_i the random variable Y_i is conditionally independent of all the remaining random variables (i.e., $\{X_j\}_{j \neq i}$ and $\{Y_j\}_{j \neq i}$) for all $i = 1, \ldots, n$, then

$$I(X_1, \ldots, X_n; Y_1, \ldots, Y_n) \leq \sum_{i=1}^{n} I(X_i; Y_i).$$

(Hint: (i) Recall that X and Y are independent if and only if $I(X;Y) = 0$, and conditionally independent given Z if and only if $I(X;Y|Z) = 0$.

(ii) In addition, the first part uses the general chain rule for entropy, as well as $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$ which is also the chain rule. The second part uses the expansion $I(\mathbf{X};\mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})$, the general chain rule for (conditional) entropy, and the suba-additivity of entropy.)

14. [**Conditional vs. Unconditional Mutual Information**] We know that conditioning reduces entropy: $H(X|Z) \leq H(X)$. Show that the same might not necessarily hold for mutual information, i.e., show that there exists a joint distribution P_{XYZ} such that $I(X;Y|Z) > I(X;Y)$.

Hints for Part II

8. Write down the joint distribution, then the relevant marginal/conditional distributions, then compute $I(U;L)$.

9. Write (conditional) mutual information in terms of two (conditional) entropies, then apply the (conditional) chain rule to each one.

10. Same idea as the previous one, but now use the general (conditional) chain rule for entropy.

11. In (a), after applying the chain rule in two ways, you will have an equality with four mutual information terms (two on each side). Argue that one is zero, and note that other one is ≥ 0. In (b), consider when the ≥ 0 inequality just mentioned holds with equality. In (c) use the same ideas again, but with the roles of X and Z swapped.

12. (a) follows easily from the expression with two conditional entropies, (b) uses symmetry of both $I(X;Y)$ and $H(X,Y)$, and for (c) see the hint in the question.

13. See the hints in the questions.

14. Think about conditional vs. unconditional independence, and recall the connection between independence and zero mutual information.