CS3236: Solutions to 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 $H_2(\cdot)$.

   Solution. *Clearly $H(X) = H_2(p)$ because $H_2(\cdot)$ is the entropy of a Bernoulli random variable by definition (see the lecture). We similarly have $H(Y) = H_2(q)$ where $q = \mathbb{P}[Y = 1]$. Now, $Y = 1$ occurs if either $X = 0$ is flipped or $X = 1$ is not flipped, so*

   $$q = (1-p)\delta + p(1-\delta).$$

   Since $H_2(\cdot)$ is symmetric, another acceptable answer would be $H_2(1-q)$ with $1 - q = \mathbb{P}[Y = 0]$, which can similarly be calculated as

   $$1 - q = (1-p)(1-\delta) + p\delta.$$

   (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.)*

   Solution. *Recall that $H(Y|X) = \sum_x P_X(x) H(Y|X = x)$. But $H(Y|X = x)$ is the same for either value of x, namely, it is $H_2(\delta)$ since the two values of Y occur with probabilities δ and $1-\delta$. Therefore,*

   $$H(Y|X) = H_2(\delta).$$

   For $H(X|Y)$, we compute $\mathbb{P}[X = 1|Y = 1]$ and $\mathbb{P}[X = 1|Y = 1]$ using the definition of conditional probability:

   $$\mathbb{P}[X = 1|Y = 1] = \frac{\mathbb{P}[X = 1 \cap Y = 1]}{\mathbb{P}[Y = 1]} = \frac{p(1-\delta)}{q}$$
   $$\mathbb{P}[X = 1|Y = 0] = \frac{\mathbb{P}[X = 1 \cap Y = 0]}{\mathbb{P}[Y = 0]} = \frac{p\delta}{1-q},$$

where $q = \mathbb{P}[Y = 1]$ as defined above. Therefore,

$$\begin{aligned} H(X|Y) &= \sum_y P_Y(y) H(X|Y = y) \\ &= qH_2\left(\frac{p(1-\delta)}{q}\right) + (1-q)H_2\left(\frac{p\delta}{1-q}\right). \end{aligned}$$

Again, you might end up with a different (but equivalent) expression if you use $\mathbb{P}[X = 0|Y = 1]$ and $\mathbb{P}[X = 0|Y = 0]$ instead of $\mathbb{P}[X = 1|Y = 1]$ and $\mathbb{P}[X = 1|Y = 0]$.

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 $\sum_{r=1}^{\infty} r2^{-r} = 2$)

   Solution. *The probability distribution of the random variable i is given by $\mathbb{P}[i = r] = \frac{1}{2^r}$. Now, the probability that $X = t$ is equal to $\sum_{r=1}^{\infty} \mathbb{P}[i = r]\mathbb{P}[j = (t + r)|i = r]$. Since we are tossing the fair coin independently $\mathbb{P}[j = (t + r)|i = r] = \frac{1}{2^t}$. Hence,*

$$\begin{aligned} \mathbb{P}[X = t] &= \sum_{r=1}^{\infty} \mathbb{P}[i = r]\mathbb{P}[j = (t + r)|i = r] \\ &= \sum_{r=1}^{\infty} 2^{-r}2^{-t} \\ &= 2^{-t}, \end{aligned}$$

   since $\sum_{r=1}^{\infty} 2^{-r} = 1$. Therefore,

$$\begin{aligned} H(X) &= \sum_{r=1}^{\infty} \mathbb{P}[X = r] \log_2 \frac{1}{\mathbb{P}[X = r]} \\ &= \sum_{r=1}^{\infty} r2^{-r} \\ &= 2 \end{aligned}$$

   by the second identity in the hint.

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 $p_1, \ldots, p_n$ in the range $[0, 1]$, the following holds:

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

   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:

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

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

Solution. *As hinted, let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameters $p_1, \ldots, p_n$. Let the index J be uniformly random on $\{1, \ldots, n\}$, and let X be the corresponding random variable X_J.*

The probability that $X = 1$ is $\frac{1}{n}\sum_{j=1}^{n} p_j$ (each i is selected with probability $\frac{1}{n}$ and then X_j is 1 with probability p_j), so we have

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

In addition, the definition of conditional entropy gives

$$\begin{aligned} H(X|J) &= \frac{1}{n}\sum_{j=1}^{n} H(X|J=j) \\ &= \frac{1}{n}\sum_{j=1}^{n} H_2(p_j), \end{aligned}$$

since once $J = j$ is selected the only uncertainty in X is in the probability-p_j event that $X_j = 1$.

Substituting the above findings into $H(X|J) \le H(X)$ completes the proof.

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 $H(p_1, \ldots, p_n)$ 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)
   Solution. *A "direct" proof that ignores the hint:*

$$\begin{aligned} H(P) &= \sum_{i=1}^{n} p_i \log_2\left(\frac{1}{p_i}\right) \\ &= p_1 \log_2\left(\frac{1}{p_1}\right) + \sum_{i=2}^{n} p_i \log_2\left(\frac{1}{p_i}\right) \\ &= p_1 \log_2\left(\frac{1}{p_1}\right) + (1-p_1)\sum_{i=2}^{n} \frac{p_i}{1-p_1} \log_2\left(\frac{1/(1-p_1)}{p_i/(1-p_1)}\right) \\ &= p_1 \log_2\left(\frac{1}{p_1}\right) + (1-p_1)\sum_{i=2}^{n} \frac{p_i}{1-p_1} \left[\log_2\left(\frac{1}{1-p_1}\right) + \log_2\left(\frac{1}{p_i/(1-p_1)}\right)\right] \\ &= p_1 \log_2\left(\frac{1}{p_1}\right) + (1-p_1)\log_2\left(\frac{1}{1-p_1}\right) + (1-p_1)\sum_{i=2}^{n} \frac{p_i}{1-p_1} \log_2\left(\frac{1}{p_i/(1-p_1)}\right) \\ &= H(p_1, 1-p_1) + (1-p_1)\, H\left(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, \ldots, \frac{p_n}{1-p_1}\right). \end{aligned}$$

A more elegant proof following the hint: Let $X \sim P$, and let $Y = \mathbf{1}\{X = 1\}$ be one if $X = 1$ and zero otherwise. By the chain rule, we have

$$\begin{aligned} H(X,Y) &= H(Y) + H(X|Y) \\ &\overset{(a)}{=} H(p_1, 1-p_1) + \mathbb{P}[Y=0]H(X|Y=0) + \mathbb{P}[Y=1]H(X|Y=1) \\ &\overset{(b)}{=} H(p_1, 1-p_1) + (1-p_1)H(X|Y=0) \\ &\overset{(c)}{=} H(p_1, 1-p_1) + (1-p_1)\, H\left(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, \ldots, \frac{p_n}{1-p_1}\right), \end{aligned}$$

where (a) uses the fact that Y is binary and the definition of conditional entropy, (b) uses $\mathbb{P}[Y = 0] = 1 - p_1$ and $H(X|Y = 1) = 0$ (given $Y = 1$, there is no uncertainty in X), and (c) uses the definition of conditional probability $\mathbb{P}[X = x|Y = 0] = \frac{\mathbb{P}[X=x \cap Y=0]}{\mathbb{P}[Y=0]} = \frac{\mathbb{P}[X=x]}{1-p}$ (the last equality holds for $x \neq 1$, since then $X = x$ means the event $Y = 0$ is redundant).
By applying the chain rule with X and Y reversed, we also have

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

since there is no uncertainty in Y given X (i.e., it is a deterministic function of X). Combining the previous two equations gives the desired result.

(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 $X \in \{1, 2, 3, \ldots\}$, the number of flips?

Solution. *From decomposability of entropy, $H(X)$ is summation of the binary entropy of the first outcome and $\frac{1}{3}$ the entropy of the remaining outcomes. The probability distribution for the remaining outcomes looks just like it did before we made the first toss. Thus,*

$$H(X) = H\left(\frac{2}{3}, \frac{1}{3}\right) + \frac{1}{3}\, H(X)\,.$$

Rearranging, $\frac{2}{3}\, H(X) = H\left(\frac{2}{3}, \frac{1}{3}\right)$, which implies that $H(X) = \frac{3}{2} H\left(\frac{2}{3}, \frac{1}{3}\right) \approx 1.3774$.

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.)*

Solution. *Let $Y = f(X)$. Then, $P_Y(y) = \sum_{x\,:\,y=f(x)} P_X(x)$. We observe for any y that*

$$\sum_{x\,:\,y=f(x)} P_X(x) \log_2 P_X(x) \;\leq\; \sum_{x\,:\,y=f(x)} P_X(x) \log_2 P_Y(y) \;=\; P_Y(y) \log_2 P_Y(y)\,.$$

where the inequality uses the fact that the $P_Y(y)$ value is the sum of all the $P_X(x)$ values in S_y, so its probability must be at least as high as any such individual $P_X(x)$ value.

Therefore,

$$\begin{aligned} H(X) &= \sum_x P_X(x) \log_2 \frac{1}{P_X(x)} \\ &= \sum_y \sum_{x\,:\,y=f(x)} P_X(x) \log_2 \frac{1}{P_X(x)} \\ &\geq \sum_y P_Y(y) \log_2 \frac{1}{P_Y(y)} \\ &= H(Y), \end{aligned}$$

where the inequality uses the first display equation above (after multiplying both sides by -1). Hence the inequality $H(X) \geq H(Y)$ is proved.

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

Solution. *Let X be the random variable that takes real values from the set $\mathcal{X} = \{-1, 1, 2\}$ with probabilities $1/4, 1/4, 1/2$ respectively. Let $Y = f(X^2)$. One can easily verify that $H(X) > H(Y)$ since $H(X) = 2$ and $H(Y) = 1$.*

In general, $H(X) = H(Y)$ for any one-to-one function $Y = f(X)$.

(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)

Solution. *This is a special case of part (a) with $X = (X_1, X_2)$, since $X_1 + X_2$ is a deterministic function of (X_1, X_2).*

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.

Solution. *Notice that in two uses of the balance – each of which reads one out of the 3 outcomes – the number of conceivable outcomes is $3^2 = 9$, whereas the number of possible states in the problem is 10: the odd ball could be any one of ten balls. Hence it cannot be solved by using just 2 uses of the balance.*

An alternative "more information-theoretic" solution. Let X be a uniformly random index specifying which of the 10 balls is heavy (under a uniform prior), and let $\mathbf{Y}$ be the sequence *of test outcomes (each entry equals $+1$, 0, or -1). For the outcomes to uniquely determine the heavy ball, we need $H(X|\mathbf{Y}) = 0$. By the chain rule, $H(X, \mathbf{Y}) = H(\mathbf{Y}) + H(\mathbf{Y}|X) = H(\mathbf{Y})$. But the chain rule in the opposite direction gives $H(X, \mathbf{Y}) \leq H(X) = \log_2 10$, so overall we require $H(\mathbf{Y}) \geq \log_2 10$. With 2 weighings, $\mathbf{Y}$ can only take at most 9 different values so $H(\mathbf{Y}) \leq \log_2 9$, and the required inequality $H(\mathbf{Y}) \geq \log_2 10$ cannot hold.*

(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.)

Solution. *We will take the approach maximizing the entropy (information) of outcomes, via the following strategy: Weigh the first N_{left} balls against the next N_{right}, and leave the remaining N_{table} balls on the table. Here N_{left}, N_{right}, N_{table} are selected to make the entropy $H(Y)$ of the outcome as high as possible (see below). The result of the weighing lets us rule out 2 of the 3 groups of balls, and after that we recursively apply this procedure to the remaining group.*

- *Note: Arguably it is not maximizing information $H(Y)$ that is most important, but minimizing remaining uncertainty $H(X|Y)$, where X denotes the index of the heavy ball. In fact, in this particular problem the two are equivalent! To see this, use the chain rule (twice) to write*

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

But $H(Y|X) = 0$ because the weighing outcomes are deterministic, and $H(X)$ is something that we have no control over. So maximizing $H(Y)$ is the same as minimizing $H(X|Y)$.

First Weighing :

For first weighting, If we choose 2 balls against 2 balls or 3 balls against 3 balls (and so on), then we will only get the outcome 1 since the odd ball is only slightly heavier or lighter than the rest. More specifically, to have any hope of gaining information we have to choose one of the following options:

- *1 ball against 2: This gives outcome probabilities 1/10, 2/10, 7/10.*
- *2 balls against 4: This gives outcome probabilities 2/10, 4/10, 4/10.*
- *3 balls against 6: This gives outcome probabilities 3/10, 6/10, 1/10.*

Note that these probabilities are calculated assuming the specified number of balls are chosen uniformly at random.

Among these options, the maximum information is gained (on average) by choosing 2 balls against 4 balls, since $H(2/10, 4/10, 4/10) = 1.52$ is higher than $H(3/10, 6/10, 1/10) = 1.29$ or $H(1/10, 2/10, 7/10) = 1.16$.

After First Weighing - Outcome (i) :

If we get outcome (i) after the first weighing, then we know that either one among balls 1,2 is heavier. Now, weigh ball 1 against balls 3,4. If the outcome is (i) again, then ball 1 is heavier. outcome (ii) is impossible and outcome (iii) implies 2 is heavier.

After First Weighing - Outcome (ii) :

If we get outcome (ii) after the first weighing, then we know that one among balls 3,4,5,6 is heavier and rest are equal. Now, weigh ball 3 against 4,5. If the outcome is (i), then ball 3 is heavier; if the outcome is (ii), then one among balls 4,5 is heavier; and outcome (iii) implies ball 6 is heavier.

We may need one more weighing to find one among 4,5 balls is the outcome is (ii) after second weighting making the total count of weighings to 3.

After First Weighing - Outcome (iii) :

If we get outcome (iii) after the first weighing, then we know that either one among 7,8,9,10 balls is heavier and rest are equal. Proceed in the same way as case (ii) for a total of up to 3 weighings.

Hence, it suffices to have 3 weighings to pick the odd ball out.

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.)

Solution. *We use Jensen's inequality. Let P, Q be probability distributions over the alphabet $\mathcal{X}$. Now $\forall x \in \mathcal{X}$, let $u = \frac{Q(x)}{P(x)}$, and define the function $f(u) = -\log(u)$. Then Jensen's inequality applied to f gives*

$$D(P||Q) = \mathbb{E}_P[f(Q(X)/P(X))] \geq f\left(\sum_x P(x)\frac{Q(x)}{P(x)}\right) = \log\left(\frac{1}{\sum_x Q(x)}\right) = 0$$

with equality if and only if $u = \frac{Q(x)}{P(x)}$ is a constant $\forall x \in \mathcal{X}$, that is, if $P = Q$.

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 $L, I(U;L)$?

Solution.

(a) The joint distribution of the upper face (U) and the lower face (L) is as follows:

		L	
	P_{UL}	black	white
U	black	1/3	1/6
	white	1/6	1/3

Using these probabilities, we can compute the following:

$$\Pr[\, L = \textit{black} \mid U = \textit{black}\,] = \frac{\Pr[\, L = \textit{black},\ U = \textit{black}\,]}{\Pr[\, U = \textit{black}\,]} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{6}} = \frac{2}{3}$$

$$\Pr[\, L = \textit{white} \mid U = \textit{black}\,] = \frac{1}{3}$$

The other variations of $P_{L|U}$ and $P_{U|L}$ can be computed similarly, and also equal to $\frac{1}{3}$ or $\frac{2}{3}$.

(b) Seeing the upper face does convey information about the color of the lower face:

- *$H(U) = H(L) = 1$ bit, since a given face is always black or white with equal probability.*
- *$I(U;L) = H(L) - H(L|U) = 1 - H_2(\frac{1}{3}) = 0.08$ bits. Note that the value $H_2(\frac{1}{3})$ comes from the fact that $H(L|U = u)$ takes that value for <u>either</u> choice of u (see part (a)).*

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

Recall the following definitions and properties.

<u>Conditional Mutual Information:</u> 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 $I(X;Y) = H(X) - H(X|Y)$.