TTIC 31200 – Information and Coding Theory – Discussion 2

Kavya Ravichandran*

17 January 2025

1 Mutual Information

Recall from lecture the definition of mutual information:

Definition 1.1. *The mutual information between two random variables X, Y is defined as:*

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\,.$$

Exercise 1.1. *Prove $I(X;Y) = 0 \iff X, Y$ are independent.*

Solution: First, let us consider the easier direction, i.e., if X, Y are independent, then $I(X;Y) = 0$. If X, Y are independent, then we know that $p(x,y) = p(x)p(y)$. Thus, for each term in the sum, we have $\frac{p(x,y)}{p(x)p(y)} = 1$, and so each term is 0. Thus, if the variables are independent, then the mutual information is 0.

In the other direction, we appeal to the strong concavity of $\log(\cdot)$. From Jensen's inequality, observe that:

$$\begin{aligned}
I(X;Y) &= -\sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} && (1)\\
&= \mathbb{E}_{x,y}\left[-\log\left(\frac{p(x)p(y)}{p(x,y)}\right)\right] && (2)\\
&\leq -\log \mathbb{E}_{x,y}\left[\frac{p(x)p(y)}{p(x,y)}\right] && (3)\\
&= -\log\left(\sum_{x,y} p(x,y)\frac{p(x)p(y)}{p(x,y)}\right) && (4)\\
&= 0\,. && (5)
\end{aligned}$$

Jensen's Inequality

So we are really interested in when Jensen's inequality is tight. Due to the strong concavity of the $\log(\cdot)$ function, since there are no places where interpolating along the linear segment will exactly coincide with the function, the only way to achieve tightness is by having all the arguments be equal, i.e., $\frac{p(x)p(y)}{p(x,y)} = \alpha$ for all x, y and some fixed α. Since we are working with probability distributions, α must be 1, and so $p(x,y) = p(x)p(y)$, the definition of independence. □

2 Sufficient Statistics

Next, let us look at an application of mutual information, sufficient statistics. Following Cover and Thomas, we have that the definition of a sufficient statistic is as follows:

Definition 2.1. *A function $T(X)$ is said to be a* sufficient statistic *relative to the family $\{f_\theta(x)\}$ if X is independent of θ given $T(X)$ for any distribution on θ, i.e., $\theta \to T(X) \to X$ forms a Markov chain.*

*Section 2 of this document draws on recitation notes developed by Max Ovsiankin when TAing a previous offering of this course.

The presence of this family $f_\theta(x)$ may be a little puzzling – it describes the set of all functions that take in x as an argument and are parametrized by θ. Since a function parameterized by θ is determined once θ is determined, all this means is that something that is a sufficient statistic for θ allows you to compute any function that requires a value of θ as a parameter.

An alternate definition (which can be thought of as a necessary and sufficient condition for the definition above) is as follows:

Definition 2.2. *$T(X)$ is sufficient relative to $\{f_\theta(x)\} \iff I(\Theta; X) = I(\Theta; T(X))$ for all distributions on Θ.*

Try proving that these definitions are equivalent.

We work through an example claiming that the number of heads in a sequence of coin flips is sufficient for heads probability parameter. Two other examples of sufficient statistics can be found in Cover and Thomas, Section 2.9.

Exercise 2.1. *The number of heads is a sufficient statistic with respect to X for a sequence of coin flips Y_i*

$$Y_i = \begin{cases} 1 & w.p.\ X \\ 0 & w.p.\ 1-X \end{cases} \qquad X = \begin{cases} p_1 & w.p.\ \frac{1}{2} \\ p_2 & w.p.\ \frac{1}{2} \end{cases}.$$

Solution: L et us begin with some intuition. Consider the probability of seeing a sequence $\{Y_i\}_{i=1}^n$ of n flips that has probability p of heads.

$$S_0 \coloneqq \{i : Y_i = 0\} \tag{6}$$
$$S_1 \coloneqq \{i : Y_i = 1\} \tag{7}$$
$$\mathbb{P}\left[\{Y_i\}_{i=1}^n = \{y_i\}_{i=1}^n\right] = \Pi_{i=1}^n p^{y_i}(1-p)^{1-y_i} \tag{8}$$
$$= \Pi_{i\in S_0}(1-p) \cdot \Pi_{i\in S_1} p \tag{9}$$
$$= (1-p)^{|S_0|} \cdot p^{|S_1|} \tag{10}$$
$$= p^{|S_1|} \cdot (1-p)^{n-|S_1|}. \tag{11}$$

The manipulation above shows us that the distribution can be written as a function of only the number of 1s in the sequence / the number of heads regardless of the value taken by p, suggesting it is *sufficient* to capture desired information about the distribution.

In order to show this formally, we would need to argue that once we condition on $|S_1| = k$, X and $\{Y_i\}_{i=1}^n$ are independent.

3 KL-Divergence (aka Relative Entropy)

The last topic we covered in the discussion section was KL-Divergence. This is also known by other names – if you are trying to read more about the topic in Cover and Thomas, look for *Relative Entropy.*

Definition 3.1. *For two distributions P, Q supported on $\mathcal{X}$, we define $D(P||Q)$, the KL-Divergence or Relative Entropy, as:*

$$D(P||Q) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

This can also be rewritten as $D(P||Q) = \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log q(x) = \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{q(x)} - H(X)$ One interpretation of the relative entropy is how much is "wasted" by encoding elements sampled from distribution P by using an encoding based on distribution Q. We can start by reinterpreting mutual information as the relative entropy between two distributions:

Exercise 3.1. *Write $I(X;Y)$ in terms of the KL Divergence.*

Solution: W e can pick the two distributions as follows:

$$I(X;Y) = D\Big(p(x,y) \Big|\Big| p(x)p(y) \Big) .$$

This tells us that we can interpret the mutual information as the loss when we encode a joint distribution as the product distribution of the marginal, i.e., by assuming that the variables are independent.

In the rest of the discussion, we had some practice with computing and interpreting the KL Divergence between distributions. For this, consider the following distribution:

$$p_0 = 0.26 \qquad p_1 = 0.39 \qquad p_2 = 0.24 \qquad p_3 = 0.08 \qquad p_4 = 0.02 \qquad p_5 = 0.01$$

For ease of computation, here are some useful numbers:

$$H(P) = 2 \qquad \log 6 = 2.585$$

Exercise 3.2. *Compute $D(P||U)$, where U is the uniform distribution over the same support as P.*

Solution: W e may start by writing out the KL Divergence and then plugging in the relevant values:

$$\begin{aligned} D(P||U) &= \sum_{i=0}^{5} p_i \log \frac{p_i}{u_i} && (12)\\ &= \sum_{i=0}^{5} p_i \log p_i - \sum_{i=0}^{5} p_i \log u_i && (13)\\ &= -H(P) - \log \frac{1}{6} && (14)\\ &= \log 6 - H(P) = 0.585 . && (15) \end{aligned}$$

Exercise 3.3. *Compute $D(U||P)$, where U is the uniform distribution over the same support as P.*

Solution: A s before, we may start by writing out the KL Divergence and then plugging in the relevant values:

$$\begin{aligned} D(U||P) &= \sum_{i=0}^{5} u_i \log \frac{u_i}{p_i} && (16)\\ &= \sum_{i=0}^{5} u_i \log u_i - \sum_{i=0}^{5} u_i \log p_i && (17)\\ &= -\log 6 - \frac{1}{6} \left(\log 0.26 + \log 0.39 + \log 0.24 + \log 0.08 + \log 0.02 + \log 0.01\right) && (18)\\ &= 0.964 . && (19) \end{aligned}$$

This is a crucial point to note – the KL Divergence is *NOT* symmetric, and therefore it is *NOT* a distance metric. So it is useful to think about the KL Divergence as how much a distribution P diverges from a reference distribution Q.

Next, we consider a very fundamental problem in statistical estimation: can we find a parametric distribution that is "close" to another distribution for which we have a full description? Suppose we receive the probabilities in P via some sort of empirical setting, e.g., the number of times someone clicked on an ad shown to them 6 times. It seems like this could reasonably be described by a binomial distribution, assuming each click is Bernoulli, distributed independently and identically. Thus, let us try to figure out what the relevant Bernoulli parameter is, i.e., the probability of an individual clicking. Formally, we find the binomial distribution that minimizes the KL Divergence between it and the observed distribution, P.

Exercise 3.4. *Suppose I want to find the best binomial approximation to P as measured by the KL Divergence. What is the right parameter β when $n = 6$.*

Solution: R ecall the binomial PMF for n trials with success probability β is given by:

$$b_i \coloneqq \mathbb{P}[Y = i] = \binom{n}{i} \beta^i (1-\beta)^{n-i}.$$

Now, we first observe that for optimizing the KL Divergence, we can ignore terms that do not depend on β, i.e., do not depend on b_i. We can then identify the relevant terms of the KL Divergence:

$$D(P||B) = \sum_{i=0}^{5} p_i \log \frac{p_i}{b_i} \tag{20}$$

$$= \sum_{i=0}^{5} p_i \log p_i - \sum_{i=0}^{5} p_i \log b_i \tag{21}$$

Thus, we focus on this second term, plugging in the definition of b_i above:

$$-\sum_{i=0}^{5} p_i \log b_i = -\sum_{i=0}^{5} p_i \log \left(\binom{6}{i} \beta^i (1-\beta)^{6-i} \right) \tag{22}$$

$$= -\sum_{i=0}^{5} p_i \left(\log \binom{6}{i} + i \log \beta + (6-i) \log(1-\beta) \right) \tag{23}$$

only non-constant terms: $\rightarrow$

$$-\sum_{i=0}^{5} p_i \left(i \log \beta + (6-i) \log(1-\beta) \right). \tag{24}$$

Now, we take the derivative of the relevant terms and set equal to 0.

$$\frac{d}{d\beta} \left(-\sum_{i=0}^{5} p_i \left(i \log \beta + (6-i) \log(1-\beta) \right) \right) = 0 \tag{25}$$

$$-\sum_{i=0}^{5} \frac{p_i \, i}{\beta \ln 2} - \sum_{i=0}^{5} \frac{-p_i (6-i)}{(1-\beta) \ln 2} = 0 \tag{26}$$

$$\frac{\sum_{i=0}^{5} p_i \, i}{\beta \ln 2} = \frac{\sum_{i=0}^{5} p_i (6-i)}{(1-\beta) \ln 2} \tag{27}$$

$$\beta = \frac{\sum_{i=0}^{5} p_i \, i}{\sum_{i=0}^{5} p_i \, i + \sum_{i=0}^{5} p_i (6-i)} \tag{28}$$

$$= \frac{\sum_{i=0}^{5} (6-i)}{6}. \tag{29}$$

solving for β, we get $\beta = 0.206.. \approx 0.21$.

Finally, to check that this is indeed a minimum (and not a maximum), we must take the second derivative and check that it is non-negative. Check this on your own for thoroughness.