CS 229r Information Theory in Computer Science Feb. 5, 2019

Lecture 3

Instructor: Madhu Sudan *Scribe: Prayaag Venkat*

1 Administrative notes

1. **Scribing:** Due to the large class size, students may double or triple up on scribing for lectures. Madhu will post further instructions.
2. **Problem Set 1:** Due Friday, February 8.
3. **Office Hours:** Madhu will hold office hours after lectures, in MD 339. See Piazza for Mitali's office hours.

2 Plan and Review

In this lecture, we covered the following topics that will give us more background on information theory:

1. Conditional entropy, Divergence, Mutual Information
2. Divergence Theorem and applications

Before proceeding, we review some concepts from the previous lecture. For a random variable X, its *entropy* $H(X)$ is the average number of bits needed to convey n i.i.d. copies $X_1, \ldots, X_n$ of X in expectation. Here, we are averaging over the n copies (dividing by n) and computing the expectation over the random variables $X_1, \ldots, X_n$. We saw that if X is supported on a finite set $\Omega = [m]$ and its distribution P_X is written as $P_X = (p_1, \ldots, p_m)$ (where $p_i \geq 0$ and $\sum_{i=1}^m p_i = 1$), then we can write:

$$H(X) = \sum_{i=1}^{m} p_i \log \frac{1}{p_i} = \mathop{\mathbb{E}}_{i \sim P_X}[\log \frac{1}{p_i}].$$

We can interpret this second expression as telling us that to encode element i, we are "budgeting" $l_i^* = \log \frac{1}{p_i}$ bits. We can then ask if this choice of $\{l_i^*\}_{i=1}^m$ is the best set of encoding lengths. Is it possible that some other $\{l_i\}_{i=1}^m$, where we encode i using l_i bits, achieves a smaller expected encoding length? Problem 4 on Problem Set 1 (Kraft's Inequality) asks you to investigate what constraints one must have on $\{l_i\}_{i=1}^m$ in order to have a valid encoding. Any prefix-free encoding must satisfy $\sum_i 2^{-l_i} \leq 1$.

Given $\{l_i\}_{i=1}^m$, we can define $q_i = 2^{-l_i}$. It is easy to see that $q_i \geq 0$ and $\sum_i q_i \leq 1$ (if a corresponding prefix-free encodinng exists, by Kraft's inequality). Then the expected number of bits we need to send is $\sum_i p_i l_i = \sum_i p_i \log(\frac{1}{q_i})$. By the end of this lecture, we hope to show that:

$$\sum_i p_i \log(\frac{1}{q_i}) \geq \sum_i p_i \log(\frac{1}{p_i}).$$

This tells us that the optimal way to compress P_X is by using $\{l_i^*\}_{i=1}^m$, rather than any other $\{l_i\}_{i=1}^m$.

3 Axioms of Entropy

First, we set up some notation. X and Y are random variables supported on Ω. Their joint distribution is P_{XY}, written $(X,Y) \sim P_{XY}$, which simply means $Pr[X=\alpha, Y=\beta] = P_{XY}(\alpha,\beta)$. The marginal distribution of X is P_X, where $P_X(\alpha) = \sum_{\beta\in\Omega} = P_{XY}(\alpha,\beta)$, and similarly for Y. The conditional distribution of Y given that $X=\alpha$ is $P_{Y|X=\alpha}$, where $P_{Y|X=\alpha}(\beta) = \frac{P_{XY}(\alpha,\beta)}{P_X(\alpha)}$. Finally, we write $X \perp Y$ to denote that X, Y are independent.

Now, recall the followings axioms. By the end of the lecture, we will formally prove all of them.

1. $H(X) \leq \log|\Omega|$, with equality iff $P_X = Unif(\Omega)$.
2. $H(X,Y) = H(X) + H(Y|X)$. This is the chain rule for entropy.
3. $H(Y|X) \leq H(Y)$. This captures the intuitive fact that conditioning can only reduce entropy.

4 Conditional Entropy

Definition 1 (Conditional entropy)**.** The *conditional entropy* of Y given X is the expected entropy of the conditional random variable $Y|X$. Formally, it is defined as:

$$H(Y|X) = \mathop{\mathbb{E}}_{\alpha\sim P_X}[H(Y|X=\alpha)] = \sum_{\alpha\in\Omega} P_X(\alpha) H(Y|X=\alpha) = \sum_{\alpha,\beta\in\Omega} P_X(\alpha) P_{Y|X=\alpha}(\beta) \log\frac{P_X(\alpha)}{P_{XY}(\alpha,\beta)}.$$

Exercise 2. *Given this definition of conditional entropy, prove Axiom 2.*

Proof. To do this, just expand out definitions:

$$\begin{aligned}
H(Y|X) &= \sum_{\alpha,\beta\in\Omega} P_X(\alpha) P_{Y|X=\alpha}(\beta) \log\frac{P_X(\alpha)}{P_{XY}(\alpha,\beta)} \\
&= \sum_{\alpha,\beta\in\Omega} P_X(\alpha) P_{Y|X=\alpha}(\beta)(\log P_X(\alpha) - \log P_{XY}(\alpha,\beta)) \\
&= \sum_{\alpha\in\Omega} P_X(\alpha)\log P_X(\alpha) + \sum_{\alpha,\beta\in\Omega} P_X(\alpha) P_{Y|X=\alpha}(\beta)\log\frac{1}{P_{XY}(\alpha,\beta)} \\
&= -H(X) + H(X,Y).
\end{aligned}$$

□

Exercise 3. *Recall that $X \perp Y$ means $P_{XY}(\alpha,\beta) = P_X(\alpha)P_Y(\beta)$ for all $\alpha,\beta\in\Omega$. Prove that if $X\perp Y$, then $H(Y|X) = H(Y)$ (this is one part of Axiom 3).*

Proof. Again, we expand out definitions and use $X \perp Y$ to factor the joint probability distribution P_{XY}.

$$\begin{aligned}
H(Y|X) &= \sum_{\alpha,\beta\in\Omega} P_X(\alpha) P_{Y|X=\alpha}(\beta) \log\frac{P_X(\alpha)}{P_{XY}(\alpha,\beta)} \\
&= \sum_{\alpha,\beta\in\Omega} P_X(\alpha) P_Y(\beta) \log\frac{P_X(\alpha)}{P_X(\alpha)P_Y(\beta)} \\
&= \sum_{\alpha,\beta\in\Omega} P_X(\alpha) P_Y(\beta) \log\frac{1}{P_Y(\beta)} \\
&= \sum_{\beta\in\Omega} P_Y(\beta)\log\left(\frac{1}{P_Y(\beta)}\right)\sum_{\alpha\in\Omega} P_X(\alpha) = H(Y).
\end{aligned}$$

□

Combining these two exercises, we easily obtain the following intuitive result that entropy is multiplicative.

Corollary 4. *IF $X_1, \ldots, X_n$ are i.i.d. copies of X then $H(X_1, \ldots, X_n) = nH(X)$.*

5 Divergence

We know return to the following central inequality:

$$\sum_i p_i \log(\frac{1}{q_i}) \geq \sum_i p_i \log(\frac{1}{p_i}).$$

From this, we can prove all the inequality parts of the axioms. The main technical tool is the following.

Theorem 5 (Divergence Theorem). *Let P, Q be distributions on Ω. Then:*

$$\mathop{\mathbb{E}}_{x \sim P}[\log \frac{1}{P(x)}] \leq \mathop{\mathbb{E}}_{x \sim P}[\log \frac{1}{Q(x)}].$$

Moreover, equality is attained iff $P = Q$.

Note that in the inequality, both expectations are taken over P. First, if $P(x) = 0$, then we can just take $P(x) \log \frac{1}{P(x)}$ to be 0. Second, if $P(x) > 0$, but $Q(x) = 0$, then the right hand side of the inequality is ∞, meaning that Q was not "expecting" x to appear.

To prove this Divergence Theorem, we will make use of Jensen's Inequality.

Theorem 6 (Jensen's Inequality). *Let $f : \mathbb{R} \to \mathbb{R}$ be a concave function and Z a real-valued random variable. Then:*

$$\mathop{\mathbb{E}}_{Z}[f(Z)] \leq f(\mathop{\mathbb{E}}_{Z}[Z]).$$

Moreover, if f is strictly concave, then equality holds iff Z is deterministic (a constant).

We omit the proof; see the Wikipedia page for an explanation.

Proof of Divergence Theorem. Apply Jensen's Inequality on the function $f(x) = \log x$ (which is strictly concave) and the random variable $Z = \frac{Q(X)}{P(X)}$ where $X \sim P$. Then it follows that:

$$\mathop{\mathbb{E}}_{X \sim P}[\log \frac{Q(X)}{P(X)}] \leq \log \mathop{\mathbb{E}}_{X \sim P}[\frac{Q(X)}{P(X)}] = 0.$$

Using linearity of expectation and rearranging, we get that:

$$\mathop{\mathbb{E}}_{x \sim P}[\log \frac{1}{P(x)}] \leq \mathop{\mathbb{E}}_{x \sim P}[\log \frac{1}{Q(x)}].$$

Finally, the equality part of the theorem follows from the equality part of Jensen's Inequality. □

Revisiting the proof, we can extract the following useful definition.

Definition 7. (Kullback-Leibler Divergence) The *KL divergence* between two distributions P, Q is:

$$D(P||Q) = \mathop{\mathbb{E}}_{X \sim P}[\log \frac{P(X)}{Q(X)}].$$

Roughly, $D(P||Q)$ represents the similarity of the two distributions. It describes the average increase in bits one would need to encode $X \sim P$ under the mistaken belief that $X \sim Q$. More explicitly, the KL divergence satisfies the following nice properties:

1. $D(P||Q) \geq 0$, with equality iff $P = Q$.
2. $D(P^n||Q^n) = nD(P||Q)$, where P^n denotes the n-fold product distribution of P.

On the other hand, the KL divergence is not so well-behaved in the following ways:

1. It is not symmetric. That is, $D(P||Q) \neq D(Q||P)$ in general.
2. It does not satisfy the triangle inequality. That is, $D(P||Q) \nleq D(P||R) + D(R||Q)$ in general.
3. $D(P||Q)$ is not bounded. This occurs, for example, when $Q(x) = 0 < P(x)$ for some element $x \in \Omega$.

5.1 Applications

We will now use the Divergence Theorem to prove the remaining parts of the axioms.

Exercise 8. *Prove Axiom 1.*

Proof. To do this, we will instantiate the Divergence Theorem with $P = P_X$ and $Q = Unif(\Omega)$:

$$\begin{aligned} H(X) &= \mathop{\mathbb{E}}_{x\sim P_X}[\log \frac{1}{P_X(x)}] \\ &\leq \mathop{\mathbb{E}}_{x\sim P_X}[\log \frac{1}{Q(x)}] \\ &= \mathop{\mathbb{E}}_{x\sim P_X}[\log |\Omega|] = \log |\Omega|, \end{aligned}$$

where the inequality becomes an equality iff $P_X = Unif(\Omega)$. □

To prove Axiom 3, we will look at the divergence between P_{XY} (the joint distribution) and $P_X \times P_Y$ (the product distribution of the marginals). Note that if $X \perp Y$, the $P_{XY} = P_X \times P_Y$. From the chain rule, we know that $H(X, Y) = H(X) + H(Y|X)$. Because $P_X \times P_Y$ is a product distribution, the entropy of a random variable distributed according to it is $H(X) + H(Y)$. If we show that $H(X, Y) \leq H(X) + H(Y)$, then we may conclude that $H(Y|X) \leq H(Y)$ (which is precisely Axiom 3).

Proceeding in this way, we know $0 \leq D(P_{XY}||P_X \times P_Y)$. Rearranging as in the proof of the Divergence Theorem, we have:

$$H(X, Y) = \mathop{\mathbb{E}}_{(x,y)\sim P_{XY}}[\log \frac{1}{P_{XY}(x, y)}] \leq \mathop{\mathbb{E}}_{(x,y)\sim P_{XY}}[\log \frac{1}{P_X(x)P_Y(y)}] = H(X) + H(Y),$$

where the last step follows by expanding the logarithm of the product and collecting terms appropriately.

6 Mutual Information

Definition 9. The *mutual information* $I(Y; X)$ of two random variables X, Y represents the amount of information that X contains about Y. Formally, we define it to be $I(Y; X) = H(Y) - H(Y|X)$.

The following corollary is implied by the third axiom.

Corollary 10. $I(Y; X) \geq 0$*, with equality iff* $X \perp Y$*.*

Exercise 11. *Verify that* $I(Y; X) = I(X; Y)$*.*

Proof. Simply apply the chain rule of entropy and expand the definitions:

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

□

Exercise 12. *Let $X \sim P_X$ and $Y \sim P_Y$. Prove that $I(X;Y) = D(P_{XY}||P_X \times P_Y)$, where P_{XY} is the joint distribution and $P_X \times P_Y$ is the product distribution of X and Y.*

Proof. We will expand out the definition of KL divergence and use the fact (see previous exercise's proof) that $I(X;Y) = H(X) + H(Y) - H(X,Y)$:

$$
\begin{aligned}
D(P_{XY}||P_X \times P_Y) &= \mathop{\mathbb{E}}_{(x,y)\sim P_{XY}}\left[\log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim P_{XY}}\left[\log \frac{1}{P_X(x)}\right] + \mathop{\mathbb{E}}_{(x,y)\sim P_{XY}}\left[\log \frac{1}{P_Y(y)}\right] - \mathop{\mathbb{E}}_{(x,y)\sim P_{XY}}\left[\log \frac{1}{P_{XY}(x,y)}\right] \\
&= \mathop{\mathbb{E}}_{x\sim P_X}\left[\log \frac{1}{P_X(x)}\right] + \mathop{\mathbb{E}}_{y\sim P_Y}\left[\log \frac{1}{P_Y(y)}\right] - \mathop{\mathbb{E}}_{(x,y)\sim P_{XY}}\left[\log \frac{1}{P_{XY}(x,y)}\right] \\
&= H(X) + H(Y) - H(X,Y) = I(X;Y).
\end{aligned}
$$

□

6.1 Conditional Mutual Information

Definition 13. The *mutual information* $I(Y;X|Z)$ of two random variables X, Y conditioned on a third random variable Z represents the amount of information that $X|Z$ contains about $Y|Z$. Formally, we define it to be $I(Y;X|Z) = \mathbb{E}_{z\sim P_Z}[I(Y;X|Z=z)] = H(Y|Z) - H(Y|X,Z)$.

Similar to entropy, we have a chain rule for mutual information. If $X_1, \ldots, X_n$ are i.i.d. copies of X, then

$$I(Y;X_1,\ldots,X_n) = I(Y;X_1) + I(Y;X_2|X_1) + \ldots + I(Y;X_n|X_1,\ldots,X_{n-1}).$$

7 More Inequalities

We now state two more inequalities. We did not have time to cover the proofs in lecture, but they follow from the machinery we have developed so far.

Theorem 14 (Data Processsing Inequality)**.** *Let $X \to Y \to \hat{X}$ be Markov chain (meaning $X, \hat{X}$ are independent, conditioned on Y). Then:*

$$I(X;\hat{X}) \leq I(X;Y).$$

This inequality models the following scenario. X is a random variable we want to predict, based on observing only the random variable Y. $\hat{X}$ represents an estimate of X, based on Y. The inequality says that our estimator cannot contain more information about X than does Y.

As a special case, one can take $\hat{X} = g(Y)$, where g is some (deterministic) function. Then $I(X;g(Y)) \leq I(X;Y)$ describes a limitation on our predictor g.

As a side note, if $H(X)$ is small, then this tells us that X should be "predictable". Similarly, if $H(X|Y)$ is small, then X should be "predictable" from Y. Problem 5 of Problem Set 1 asks you to investigate this intuition and prove Fano's Inequality.