CS 229r Information Theory in Computer Science February 4, 2016

Lecture 3

Lecturer: Madhu Sudan *Scribe: Alex Wang*

1 Kraft's Inequality

Recall that if C is a prefix-free code $C : [n] \to \{0,1\}^*$ with $|C(1)| = l_1, |C(2)| = l_2, \ldots, |C(n)| = l_n$, then $\sum_i 2^{-l_i} \leq 1$. In the previous lecture, the proof of Kraft's Inequality claimed that any greedy code would do. However, consider the following example: $l_1 = m = l_2$ and $l_3 = 1$, e.g. $C(1) = 00000$ and $C(2) = 11111$. Then $C(3)$ will be forced to be a prefix. The fix is to have $l_1 \leq l_2 \leq \cdots \leq l_n$. Then if you pick $C(1), \ldots, C(i)$ greedily, the fraction of strings of length $l \geq l_i$ that are not allowed is $\leq \sum_{j=1}^{i} 2^{-l_j}$.

2 Basic Concepts in Information Theory

Recall our notation:

- X is a random variable with distribution $P_X = (P_1, \ldots, P_k)$ over a finite set $\Omega = \{1, \ldots, k\}$.
- The entropy of X is $H(X) = \sum_{x \in \Omega} P_x \log \frac{1}{P_x}$.
- The conditional entropy of X conditioned on another random variable Y with distribution P_Y is $H(X|Y) = \mathrm{Exp}_{Y \sim P_Y}[H(X|Y = y)]$

We also have the following axioms about entropy:

1. $H(X) \leq \log |\Omega|$
2. $H(U) = \log |\Omega|$ where U is the uniform distribution on Ω
3. $H(X|Y) \leq H(X)$
4. $H(X, Y) = H(X) + H(Y|X)$, also known as the "chain" rule

The last axiom is provable via calculations with the other three axioms while the second also follows from the definition of entropy. The first and third remain to be proven, and will be by the end of this lecture.

2.1 Mutual Information

Intuitively, the mutual information I(X; Y) between two random variables X and Y is the amount of information Y gives about X. Mathematically, we define this as the difference between the entropy of X and the entropy of X conditioned on knowing Y.

$$I(X;Y) \triangleq H(x) - H(X|Y)$$

Mutual information is symmetric, i.e. $I(X;Y) = I(Y;X)$, but in practical applications we are often more worried about one direction than the other even though mathematically they are the same. Its symmetric can be proven by using axiom 4:

$$H(X) + H(Y|X) = H(X,Y) = H(Y,X) = H(Y) + H(X|Y)$$

$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$$I(X;Y) = I(Y;X)$$

Examples:

1. If $X \perp Y, I(X;Y) = 0$ because $H(X) = H(X|Y)$
2. If $X = Y, I(X;Y) = H(X)$
3. If $X = f(Y), I(X;Y) = H(X)$ where f is a deterministic function, not necessarily invertible
4. If $Y = g(X), I(X;Y) = H(Y)$

2.2 Conditional Information

Conditional information is the expectation of the mutual information between two variables X and Y conditional on knowing the value of a third variable Z:

$$I(X;Y|Z) = \mathrm{Exp}_{Z \sim P_Z}[I(X|_{Z=z}; Y|_{Z=z})]$$

Mutual information has a chain rule involving conditional mutual information:

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

Examples:

1. Consider random variables X and Y with $X \perp Y$ and both distributed uniformly over $\{0,1\}$. Let $Z = X \oplus Y$. Then $I(X;Y) = 0$ but $I(X;Y|Z) - 1$ because we can determine X from Y and Z. In this example, conditioning on Z increases the amount of information gained by Y.
2. Now let $Z = X = Y$. Then $I(X;Y) = 1$ but $I(X;Y|Z)$ because Z already reveals everything about X; nothing is gained from Y. Thus, conditioning on Z decreases the amount of information gained by Y.

2.3 Concavity of Entropy

Consider the tuple of random variables $(X,Y) \sim P_{x,y}$ and $B \perp X, Y$ which has a distribution on $(0,1)$. Then if we let $Z = B \cdot X + (1 - B) \cdot Y$, entropy $H(Z)$ is concave if

$$H(Z) \geq \Pr[B = 0]H(Y) + \Pr[B = 1]H(x)$$

Proof: the right hand side is $H(Z|B)$, so by axiom 3 the above inequality holds.

3 Basic Inequalities in Information Theory

3.1 Data Processing Inequality

Note: we use the following notation for Markov chains:

$$X \to Y \to Z$$

In the above Markov chain, X is independent of Z conditional on Y. The data processing inequality simply tells us that the mutual information between X and Z is upper bounded by the mutual information between X and Y:

$$I(X;Z) \leq I(X;Y)$$

3.2 Fano's Inequality

We again have a Markov chain:

$$X \to Y \to \hat{X}$$

We could interpret the above, for example, as a true value X being transformed by noise into Y, and then decoded into $\hat{X}$. Then We have a probability of error P_e of the correctly recovering the true X: $P_e \triangleq Pr[X \neq \hat{X}]$. Fano's inequality gives us a lower bound on P_e:

$$P_e \geq \frac{H(X|Y) - 1}{\log |\Omega|}$$

Proof: Note that proving the above is equivalent to proving

$$P_e \log |\Omega| \geq H(X|Y) - 1$$

$$1 + P_e \log |\Omega| \geq H(X|Y)$$

By the data processing inequality, we know that $I(X; \hat{X}) \leq I(X; Y)$ so $H(X|\hat{X}) \geq H(X|Y)$. Then it suffices to prove that

$$1 + P_e \log |\Omega| \geq H(X|\hat{X})$$

Define the random variable E to be 1 if $\hat{X} \neq X$, 0 otherwise, then we will actually prove something slightly better, namely that

$$H(E) + P_e \log |\Omega| \geq H(X|\hat{X})$$

We start with the following:

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

where the last line follows since E is known if both X and $\hat{X}$ are known, so $H(E|X, \hat{X}) = 0$. Also,

$$H(E, X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X})$$

We know by axiom 3 that $H(E|\hat{X}) \leq H(E)$. For $H(X|E, \hat{X})$ if $E = 0$, which happens with probability $1 - P_e$, then since we condition on $\hat{X}$, we can determine X. In this case $H(X|E, \hat{X}) = 0$. With probability P_e, $E = 1$ so we have $H(X|E, \hat{X})$, which by our axioms we know is $\leq H(X) \leq \log |\Omega|$. So $H(X|E, \hat{X})$ is bounded by $P_e \log \Omega$. Combining all of this, we have

$$H(X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}) \leq H(E) + P_e \log |\Omega|$$

$$H(X|\hat{X}) \leq H(E) + P_e \log |\Omega|$$

3.3 Divergence Inequality

Given two distributions P and Q on Ω, we define the divergence $D(P||Q)$ of P and Q as

$$D(P||Q) \triangleq \text{Exp}_{X \sim P}[- \log \frac{Q(x)}{P(x)}]$$

Divergence lacks many useful properties. For example, divergence is not symmetric ($D(P||Q) \neq D(Q||P)$ and does not respect the triangle inequality. Because of this, divergence should not be taken to mean the "distance" between two distributions. However, divergence is useful in situations in which we have n draws from P and Q because it scales nicely in n, i.e. $D(P^n||Q^n) = n \cdot D(P||Q)$.

The divergence inequality tells us that this quantity is nonnegative, i.e.

$$\text{Exp}_{X \sim P}[- \log \frac{Q(x)}{P(x)}] \geq 0$$

Equivalently,

$$\text{Exp}_{X \sim P}[- \log Q(x)] \geq \text{Exp}_{X_P}[- \log P(x)]$$

Notice that the right hand term is the entropy of P and the left side is similar. Then we can understand the divergence inequality as saying that the entropy of an optimal compression algorithm, i.e. that makes use of the true underlying distribution, is better than the entropy of a suboptimal compression, i.e. based on a different distribution.

3.3.1 Jensen's Inequality

In order to prove the divergence inequality, we need to make use of Jensen's inequality, which we state here.

Consider a convex function $f : \mathbb{R} \to \mathbb{R}$ and a random variable X with distribution P over $\mathbb{R}$ such that $\text{Exp}_{X\sim P}[f(x)]$ and $f(\text{Exp}_{X\sim P}[x])$ are finite. Then Jensen's inequality says that

$$\text{Exp}[f(x)] \geq f(\text{Exp}[x])$$

3.3.2 Proof of Divergence Inequality

By Jensen's inequality and the convexity of $-\log(x)$, we have that

$$\text{Exp}_{X\sim P}[-\log \frac{Q(x)}{P(x)}] \geq -\log(\text{Exp}_{X\sim P}[\frac{Q(x)}{P(x)})$$

We will manipulate the right hand side:

$$-\log(\text{Exp}_{X\sim P}[\frac{Q(x)}{P(x)}) = -\log(\sum_x P(x)\frac{Q(x)}{P(x)}$$

$$= -\log(\sum_x Q(x))$$

$$= -\log(1) = 0$$

where the last line follows because $Q(x)$ is a distribution.

3.3.3 Applications of the Divergence Inequality

We can use the divergence inequality to prove axioms 1 and 3.

3.3.4 Proof of First Axiom

We have a random variable X has distribution P on Ω. The entropy of X is upper-bounded:

$$H(X) \leq \log|\Omega|$$

By the divergence inequality we know that

$$\text{Exp}_{X\sim P}[\log \frac{1}{P(x)}] \leq \text{Exp}_{X\sim P}[\log \frac{1}{Q(x)}]$$

Letting $Q(x)$ be the uniform distribution over Ω,

$$\text{Exp}_{X\sim P}[\log \frac{1}{P(x)}] \leq \text{Exp}_{X\sim P}[\log \Omega]$$

$$H(X) \leq \log \Omega$$

3.3.5 Proof of Third Axiom

Recall the axiom 3:

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

We know by the axiom 4 that $H(X,Y) = H(X) + H(Y|X)$. Then it suffices to prove

$$H(X|Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$$

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

We begin by expanding the definition of entropy:

$$\text{Exp}_{X,Y\sim P_{x,y}}[\log \frac{1}{P_{xy}(x,y)}] \leq \text{Exp}_{X,Y\sim P_{x,y}}[\log \frac{1}{P_x(x)}] + \text{Exp}_{X,Y\sim P_{x,y}}[\log \frac{1}{P_y(y)}]$$

Where the right side says that for both X and Y, we draw from $P_{x,y}$ and only keep X or Y respectively. Our goal is to transform the right side into the form $\text{Exp}_{X,Y\sim P_{x,y}}[\log \frac{1}{Q(x,y)}]$ to be able to apply the divergence inequality. By the linearity of expectation we have

$$= \text{Exp}_{X,Y\sim P_{x,y}}[\log \frac{1}{P_x(x)} + \log \frac{1}{P_y(y)}]$$

$$= \text{Exp}_{X,Y\sim P_{x,y}}[\log \frac{1}{P_x(x)P_y(y)}]$$

$$= \text{Exp}_{X,Y\sim P_{x,y}}[\log \frac{1}{Q(X,Y)}]$$

where $Q(x,y) = P_x(x)P_y(y)$. By the divergence inequality we know this is true.