CS 15-859: Algorithms for Big Data Fall 2017

Lecture 9-1 — 11/2/2017

Prof. David Woodruff *Scribe: Roie Levin*

1 Information Theory

1.1 Core Definitions

Let us consider distributions p over a finite support of size n. We write p as:

$$p = (p_1, p_2, \cdots, p_n)$$

where $p_i \in [0, 1]$ and $\sum_i p_i = 1$. We say X is a random variable with distribution p if $\mathbb{P}(X = i) = p_i$.

Definition (Shannon Entropy)**.** The Entropy of a random variable X, written $H(X)$, is:

$$\begin{aligned} H(X) &= \sum_i p_i \log_2 \frac{1}{p_i} \\ &= \mathbb{E}_p \left[\log_2 \frac{1}{p_i} \right] \end{aligned}$$

By convention, if $p_i = 0$ we define $0 \log_2 \frac{1}{0} = 0$.

Why does this choice of function make sense?

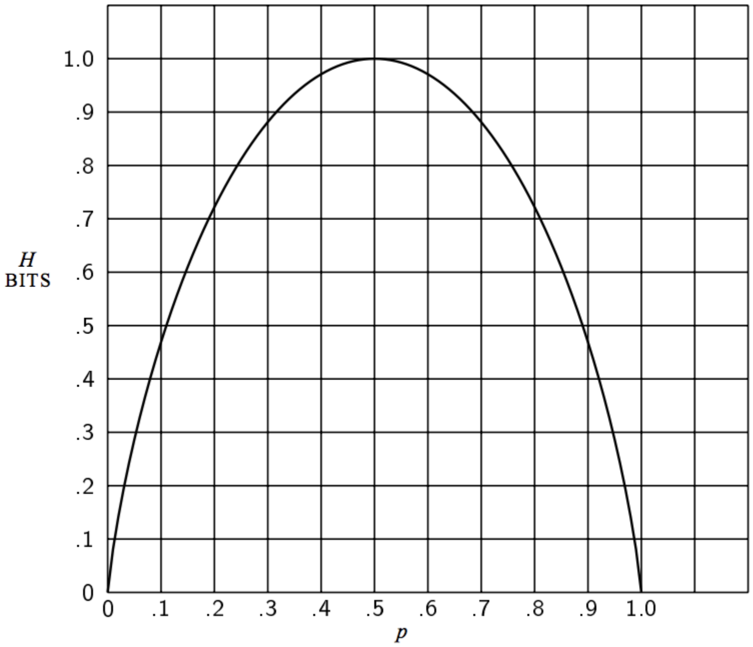
For one, note that $H(X) \leq \log_2 n$. Equality is achieved when $p_i = \frac{1}{n}$ for all i, in other words when the distribution is uniform. One can check this by solving the constrained entropy maximization problem on the probability simplex with Lagrange multipliers. Furthermore, entropy is always nonnegative, and is equal to 0 when X is just a constant with no randomness. Thus, the entropy measures the "uncertainty" of X, or informally how close it is to being uniform on the domain.

For some more intuition, suppose we wanted a function that measures the surprise of an event. The lower the probability, the more surprised we should be to see it, so the function should be decreasing with probability. Furthermore, if we want surprise to be continuous, as well as additive over independent events, then $\log \frac{1}{p}$ is a good choice. Thus intuitively Shannon Entropy is measuring expected surprise.

In the special case that X is a binary random variable B with bias p:

$$H(B) = p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p}$$

which is a symmetric function in p called the binary entropy function and is sometimes written as $H(p)$.



Definition (Conditional Entropy)**.** Given two random variables X and Y, the Conditional Entropy of X given Y, written $H(X|Y)$, is:

$$\begin{aligned} H(X|Y) &= \sum_y H(X|Y=y) \cdot \mathbb{P}\left(Y=y\right) \\ &= \mathbb{E}_y \left[H(X|Y=y)\right] \end{aligned}$$

In words, this is the average over values Y can take of the entropy of the variable X conditioned on Y taking that value.

Intuitively, this is the average uncertainty in X after we observe Y. In the special case that $Y = X$, $H(X|X) = 0$ which captures the fact that we learn everything about X from X. In the special case that X and Y are independent, $H(X|Y) = H(X)$, which captures that we learn nothing about X from Y.

Definition (Joint Entropy)**.** Given two random variables X and Y, the Joint Entropy of X and Y, written $H(X,Y)$, is:

$$H(X,Y) = \sum_{x,y} \mathbb{P}\left(X = x \wedge Y = y\right) \log_2 \frac{1}{\mathbb{P}\left(X = x \wedge Y = y\right)}$$

This is exactly the entropy of the random variable that is the tuple (X,Y).

1.2 Some Useful Facts

Theorem 1 (Chain Rule)**.** *Let X and Y be random variables. Then:*

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

Proof. Like many proofs in information theory, this fact simply come from expanding out definitions:

$$\begin{aligned}
H(X,Y) &= \sum_{x,y} \mathbb{P}\left(X = x \wedge Y = y\right) \cdot \log_2 \frac{1}{\mathbb{P}\left(X = x \wedge Y = y\right)} \\
&= \sum_{x,y} \mathbb{P}\left(X = x\right) \mathbb{P}\left(Y = y | X = x\right) \cdot \log_2 \frac{1}{\mathbb{P}\left(X = x\right) \mathbb{P}\left(Y = y | X = x\right)} \\
&= \sum_{x,y} \mathbb{P}\left(X = x\right) \mathbb{P}\left(Y = y | X = x\right) \cdot \left[\log_2 \frac{1}{\mathbb{P}\left(X = x\right)} + \frac{1}{\mathbb{P}\left(Y = y | X = x\right)}\right] \\
&= \sum_{x} \mathbb{P}\left(X = x\right) \log_2 \frac{1}{\mathbb{P}\left(X = x\right)} \underbrace{\sum_{y} \mathbb{P}\left(Y = y | X = x\right)}_{=1} \\
&\quad + \sum_{x} \mathbb{P}\left(X = x\right) \left[\sum_{y} \mathbb{P}\left(Y = y | X = x\right) \frac{1}{\mathbb{P}\left(Y = y | X = x\right)}\right] \\
&= H(X) + H(Y|X)
\end{aligned}$$

■

Theorem 2 (Conditioning Cannot increase Entropy)**.** *Let X and Y be random variables. Then:*

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

Intuitively, this means that learning information about another variable Y can only decrease the uncertainty of X.

For this proof we will need Jensen's inequality.

Fact 1 (Jensen's Inequality)**.** If f is a continuous and concave function, and $p_1, \cdots, p_n$ are nonnegative reals summing to 1, then for any $x = x_1, \cdots x_n$:

$$\sum_{i=1}^{n} p_i f(x_i) \leq f\left(\sum_{i=1}^{n} p_i x_i\right)$$

If we treat $(p_1, \cdots, p_n)$ as a distribution p, and $f(x)$ is the vector obtained by applying f coordinatewise to x then we can write the inequality as:

$$\mathbb{E}_p\left[f(x)\right] \leq f(\mathbb{E}_p\left[x\right])$$

Recall that f is said to be concave if $f\left(\frac{a+b}{2}\right) \geq \frac{f(a)}{2} + \frac{f(b)}{2}$, and $f(x) = \log_2 x$ is concave.

Proof of Theorem 2. Again by expanding definitions:

$$\begin{aligned}
H(X|Y) - H(X) &= \sum_{x,y} \mathbb{P}\left(Y = y\right) \mathbb{P}\left(X = x|Y = y\right) \cdot \log_2 \frac{1}{\mathbb{P}\left(X = x|Y = y\right)} \\
&\quad - \sum_{x} \mathbb{P}\left(X = x\right) \log_2 \frac{1}{\mathbb{P}\left(X = x\right)} \underbrace{\left[\sum_{y} \mathbb{P}\left(Y = y|X = x\right)\right]}_{=1} \\
&= \sum_{x,y} \mathbb{P}\left(X = x \wedge Y = y\right) \cdot \log_2 \frac{\mathbb{P}\left(X = x\right)}{\mathbb{P}\left(X = x|Y = y\right)} \\
&= \sum_{x,y} \mathbb{P}\left(X = x \wedge Y = y\right) \cdot \log_2 \frac{\mathbb{P}\left(X = x\right) \mathbb{P}\left(Y = y\right)}{\mathbb{P}\left(X = x \wedge Y = y\right)} \qquad (*) \\
&\leq \log_2 \left[\sum_{x,y} \mathbb{P}\left(X = x \wedge Y = y\right) \frac{\mathbb{P}\left(X = x\right) \mathbb{P}\left(Y = y\right)}{\mathbb{P}\left(X = x \wedge Y = y\right)}\right] \\
&\qquad \text{(applying Jensen's with } f = \log_2\text{)} \\
&= \log_2 1 \\
&= 0
\end{aligned}$$

■

Note that if X and Y are independent, then $\mathbb{P}\left(X = x \wedge Y = y\right) = \mathbb{P}\left(X = x\right) \mathbb{P}\left(Y = y\right)$, which means $(*)$ is exactly 0 and $H(X|Y) = H(X)$.

1.3 Mutual Information

Motivated by Theorem 2, we now give a name to the slack in Jensen's inequality above.

Definition (Mutual Information)**.** The Mutual Information of two random variables X and Y, written $I(X;Y)$, is:

$$\begin{aligned}
I(X;I) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= I(Y;X)
\end{aligned}$$

Intuitively this is a notion of how much information Y reveals about X. As a check that this makes sense, in the case that $Y = X$, $I(X;X) = H(X) - H(X|X) = H(X)$. In the case that X and Y are independent, as noted above $I(X;Y) = H(X) - H(X|Y) = 0$.

Definition (Conditional Mutual Information)**.** The Conditional Mutual Information of two random variables X and Y given random variable Z, written $I(X;Y|Z)$, is:

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

Q: Is it always the case that $I(X;Y|Z) \geq I(X;Y)$ or $I(X;Y|Z) \leq I(X;Y)$?

A: No! They can both be false.

Claim 1. For certain random variables X, Y and Z, $I(X;Y|Z) \leq I(X;Y)$.

Proof. Consider $X = Y = Z$. Then $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = 0$ ■

Intuitively, Y only reveals information about X that Z already reveals.

Claim 2. For certain random variables X, Y and Z, $I(X;Y|Z) \geq I(X;Y)$.

Proof. Consider X, Y independent uniform on $\{0,1\}$. Let $Z = X + Y \mod 2$. Then $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = 1 - 0 = 1$. On the other hand $I(X;Y) = H(X) - H(X|Y) = 1 - 1 = 0$ ■

Intuitively, Y only reveals useful information about X after we observe Z. There is also a chain rule for mutual information:

Theorem 3 (Chain Rule for Mutual Information)**.** *Let X, Y and Z be random variables. Then:*

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X)$$

Proof. Expanding definitions and using the chain rule for entropy:

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

■

By induction, this chain rule also implies that for a set of random variables $X_1, X_2, \cdots X_n, Z$:

$$I(X_1, X_2, \cdots X_n; Z) = \sum_i I(X_i; Z|X_1, \cdots X_{i-1})$$

1.4 Fano's Inequality

We now come to an important theorem that we will use later when proving communication lower bounds. Recall that $A \to B \to C$ is a Markov chain if once we condition on B, A and C are independent. "Past and future are independent given the present".

Theorem 4 (Fano's Inequality)**.** *Given a Markov chain $X \to Y \to X'$, and suppose $P_e = \mathbb{P}(X \neq X')$. Then:*

$$H(X|Y) \leq H(P_e) + P_e \log_2(|X| - 1)$$

where $|X|$ denotes the size of the support of X.

We can interpret X as a message sent across a noisy channel, Y as the signal received on the other end, and X' as an estimate of X reconstructed only from Y.

Before we prove Fano's inequality, we prove an intermediate statement:

Theorem 5 (Data Processing Inequality). *If is $X \to Y \to Z$ a Markov chain, then:*

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

Intuitively, if Y is a noisy estimator for X, then no amount of clever data combination that only uses Y can give more information about X than Y itself.

Proof. Using the chain rule for mutual information:

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

Thus it suffices to show that $I(X;Z|Y) = 0$.

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

But X and Z are independent given Y, so $H(X|Y,Z) = H(X|Y)$, and thus the claim is proved. ■

Note, the Data Processing Inequality means that $H(X|Y) \leq H(X|Z)$ because:

$$\underbrace{I(X;Y)}_{H(X)-H(X|Y)} \geq \underbrace{I(X;Z)}_{H(X)-H(X|Z)}$$

Proof of Fano's Inequality. We use the standard information theory trick of expanding a definition in two different ways.

Let E be the random variable that is 1 if X' is not equal to X and 0 otherwise.

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

(Since conditioning does not increase entropy)

But:

$$\begin{aligned} H(X|E,X') &= \mathbb{P}\left(E=0\right) \cdot H(X|X',E=0) + \mathbb{P}\left(E=1\right) \cdot H(X|X',E=1) \\ &\leq (1-P_e) \cdot 0 + P_e \cdot \log_2(|X|-1) \end{aligned}$$

Combining the statements above:

$$H(X|X') \leq H(P_e) + P_e \log_2(|X|-1)$$

By the Data Processing Inequality:

$$H(X|Y) \leq H(X|X') \leq H(P_e) + P_e \cdot \log_2(|X|-1)$$

■

Q: When is Fano's Inequality tight?

A: Suppose the distribution p of x satisfies $p_1 \geq p_2 \geq \cdots \geq p_n$, and suppose Y is a constant. So $I(X;Y) = H(X) - H(X|Y) = 0$. Then the best predictor of X is $X = 1$. In this case $P_e = \mathbb{P}(X' \neq X) = 1 - p$. Fano's Inequality predicts:

$$H(X|Y) \leq H(p_1) + (1-p_1)\log_2(n-1)$$

But $H(X) = H(X|Y)$, and if $p_2 = p_3 = \cdots = p_n = \frac{1-p_1}{n-1}$, then the inequality is tight! In this case:

$$\begin{aligned} H(X) &= \sum_i p_i \log_2 \frac{1}{p_i} \\ &= p_1 \log_2 \frac{1}{p_1} + \sum_{i>1} \frac{1-p_1}{n-1} \log_2 \frac{n-1}{1-p_1} \\ &= p_1 \log_2 \frac{1}{p_1} + (1-p_1)\log_2 \frac{1}{1-p_1} + (1-p_1)\log_2(n-1) \\ &= H(p_1) + (1-p_1)\log_2(n-1) \end{aligned}$$

2 Distances Between Distributions

2.1 Definitions

In this section we explore various notions of distance between distributions.

For the remainder of these notes, we assume that p and q are distributions on the same support.

Definition (Total Variation Distance)**.** The Total Variation Distance between p and q, denoted $D_{TV}(p,q)$ is:

$$D_{TV}(p,q) = \frac{1}{2}\|p-q\|_1 = \frac{1}{2}\sum_i |p_i - q_i|$$

Sometimes we abuse notation and write $D_{TV}(X,Y)$ to mean $D_{TV}(p,q)$ when random variables X and Y have distribution p and q respectively.

It is not hard to see that one can also write the Total Variation Distance in the following way as well:

$$D_{TV}(p,q) = \max_{\text{event } E} |p(E) - q(E)|$$

Define $E := \{i : p_i \geq q_i\}$. Then this achieves the max of the function above, and:

$$\begin{aligned} |p(E) - q(E)| &= \sum_{i:\ p_i \geq q_i} p_i - q_i \\ &= \sum_{i:\ q_i \geq p_i} q_i - p_i \\ &= \frac{1}{2}\|p-q\|_1 && \text{(Since } \{i:\ p_i \geq q_i\} \cup \{i:\ q_i \geq p_i\} \text{ is the entire support)} \end{aligned}$$

Note that the definitions above extend to the continuous case, in which case the sums are replaced with integrals, and the max is replaced with sup.

Definition (Hellinger Distance)**.** The Hellinger Distance between p and q, denoted $h(p,q)$, is:

$$h(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$$

If we define $\sqrt{p} = (\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_n})$ and $\sqrt{q} = (\sqrt{q_1}, \sqrt{q_2}, \cdots, \sqrt{q_n})$, we can write:

$$h(p,q) = \frac{1}{\sqrt{2}} \left\|\sqrt{p} - \sqrt{q}\right\|_2$$

Note since both D_{TV} and h are norms, they automatically satisfy the triangle inequality.

Q: Why the Hellinger Distance?

A: It turns out to be useful for independent random variables, since it has the following nice product structure in that case.

Theorem 6. *Suppose X and Y are independent random variables with distributions p and q respectively.*

$$\mathbb{P}\left((X,Y) = (x,y)\right) = p(x) \cdot q(y)$$

Suppose A and B are independent random variables with distributions p' and q' respectively.

$$\mathbb{P}\left((A,B) = (a,b)\right) = p'(a) \cdot q'(b)$$

Then if h^2 is the squared Hellinger distance:

$$h^2\left((X,Y),(A,B)\right) = 1 - (1 - h^2(X,A) \cdot (1 - h^2(Y,B))$$

Proof. Let (p,q) denote the product distribution of the variables (X,Y), and (p',q') denote the product distribution of the variables (A,B). First note that:

$$\begin{aligned}
h^2(s,t) &= \frac{1}{2}\left\|\sqrt{s} - \sqrt{t}\right\|_2^2 \\
&= \frac{1}{2}\left(\underbrace{\|\sqrt{s}\|_2^2}_{=1} + \underbrace{\left\|\sqrt{t}\right\|_2^2}_{=1} - 2\left\langle\sqrt{s}, \sqrt{t}\right\rangle\right) \\
&= 1 - \left\langle\sqrt{s}, \sqrt{t}\right\rangle \\
1 - h^2(s,t) &= \left\langle\sqrt{s}, \sqrt{t}\right\rangle \qquad\qquad (*)
\end{aligned}$$

Applying $(*)$ to $s = (p,q)$, $t = (p',q')$:

$$\begin{aligned}
h^2((p,q),(p',q')) &= 1 - \left\langle\sqrt{(p,q)}, \sqrt{(p',q')}\right\rangle \\
&= 1 - \sum_{i,j}\sqrt{p_i}\sqrt{q_j}\sqrt{p'_i}\sqrt{q'_j} \\
&= 1 - \sum_i \sqrt{p_i}\sqrt{p'_i}\sum_j \sqrt{q_j}\sqrt{q'_j} \\
&= 1 - \left(1 - h^2(p,p')\right)\left(1 - h^2(q,q')\right)
\end{aligned}$$

(reapplying $(*)$ to $s = p$, $t = p'$, and then to $s = q$, $t = q'$)

■