COMPSCI 650 Applied Information Theory	Jan 21, 2016
Lecture 2	
Instructor: Arya Mazumdar	*Scribe:* [redacted]

1 Entropy

Definition: *Entropy* is a measure of uncertainty of a random variable. The entropy of a discrete random variable X with alphabet $\mathcal{X}$ is

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

When the base of the logarithm is 2, entropy is measured in bits.

Example: One can model the temperature in a city (*e.g.* Amherst) as a random variable, X. Then the entropy of X measures the uncertainty in Amherst temperature. Let Y be a random variable representing the temperature in Northampton. We know that X and Y are not independent (they are usually quite similar). Hence, when Y is given, some uncertainty about X goes away.

Example: Consider a fair die with pmf $p(1) = p(2) = ... = p(6) = 1/6$. Its entropy is

$$H(x) = -6 \cdot \frac{1}{6} \log \frac{1}{6} = \log 6$$

Maximum entropy is achieved when all outcomes of a random variable occur with equal probability. (Note: you can prove this by assigning a variable p_i to the probability of outcome i. Then, partially-differentiate the entropy function with respect to each p_i. Set the derivatives to zero and solve for the p_i's. You will see that they are equal.)

In general, for M equally-probable outcomes, the entropy is $H(X) = \log M$.

1.1 Joint Entropy

Definition: For two random variables X and Y, $x \in \mathcal{X}, y \in \mathcal{Y}$, *joint entropy* is defined as

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y)$$

where $p(x,y) = Pr[X = x, Y = y]$ is the joint *pmf* of X and Y.

1.2 Conditional Entropy

Definition: The *conditional entropy* of a random variable Y given $X = x$ is

$$H(Y|X = x) = -\sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x)$$

When a particular value of x is not given, we must average over all possible values of X:

$$\begin{aligned} H(Y|X) &= -\sum_{x \in \mathcal{X}} p(x) \left(\sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \right) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x) \end{aligned}$$

The conditional entropy of X given Y is

$$H(X|Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x|y)$$

In general, $H(X|Y) \neq H(Y|X)$.

1.3 Chain Rule for Entropy

The *Chain Rule for Entropy* states that the entropy of two random variables is the entropy of one plus the conditional entropy of the other:

$$H(X,Y) = H(X) + H(Y|X) \tag{1}$$

Proof:

$$\begin{aligned}
H(X,Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \Big(p(x)p(y|x)\Big) \\
&= -\sum_{x \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}} p(x,y) \right) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x) \\
&= -\sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

Similarly, it can also be shown that

$$H(X,Y) = H(Y) + H(X|Y) \tag{2}$$

From (1) and (2), we see that

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

$I(X;Y)$ is known as **Mutual Information**, which can be thought of as a measure of reduction in uncertainty.

Example: Consider the random variables $X \in \{0,1\}, Y \in \{0,1\}$, representing two coin tosses. Their joint distribution is shown in Table 1.

Y\X	0	1
0	1/2	1/4
1	1/8	1/8

Table 1: Joint distribution of two coin tosses.

The joint entropy of X and Y is

$$H(X,Y) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{2}{8}\log\frac{1}{8} = 1.75$$

Note that if all probabilities were equal, we would have

$$H(X,Y) = \log 4 = 2 \text{ bits, which is the maximum entropy.}$$

The individual entropies are

$$H(X) = -\frac{5}{8}\log\frac{5}{8} - \frac{3}{8}\log\frac{3}{8} \approx 0.9544$$
$$H(Y) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} \approx 0.8113$$

Question: What is the conditional entropy of X given Y?
Answer: One possible way of solving this problem is to compute the conditional distribution of X given Y, for all possible values of X and Y. However, since we have already determined the joint and individual entropies, we can instead use the Chain Rule for Entropy: $H(X,Y) = H(Y) + H(X|Y)$.

$$H(X|Y) = 1.75 - 0.8113 \approx 0.9387$$

2 Relative Entropy

Let X be a random variable with alphabet $\mathcal{X} = \{0, 1\}$. Consider the following two distributions:

$$p(0) = \frac{1}{4} \qquad\qquad q(0) = \frac{1}{2}$$
$$p(1) = \frac{3}{4} \qquad\qquad q(1) = \frac{1}{2}$$

Let r be another probability distribution, defined below:

$$r(0) = \frac{1}{8}$$
$$r(1) = \frac{7}{8}$$

Question: Is r closer to p or q?
Answer: r is closer to p, because they are both biased toward the outcome $X = 1$.

How do we measure this similarity? One way is to use **relative entropy**.

Definition: *Relative entropy*, also known as **Divergence** or **Kullback-Leibler Distance**, is defined by

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

Property 1: Relative entropy is *not symmetric*. In other words, it is not necessarily true that $D(p||q) = D(q||p)$.

Property 2: Relative entropy is always non-negative: $D(p||q) \geq 0$. Equality is achieved only if $p(x) = q(x),\ \forall\ x \in \mathcal{X}$. This property is also known as **Information Inequality**.

Property 3: Relative entropy does not satisfy the triangle inequality.

Because KL-distance does not obey the triangle inequality and is not symmetric, it is not a true metric.

Example: Consider the distributions p and q introduced earlier, where $p,\ q \in\ \{0, 1\},\ p(0) = 1/4,\ p(1) =$

$3/4,\ q(0) = 1/2,\ q(1) = 1/2.$

$$D(p||q) = \frac{1}{4}\log\left(\frac{1/4}{1/2}\right) + \frac{3}{4}\log\left(\frac{3/4}{1/2}\right) = \frac{3}{4}\log 3 - 1 \approx 0.1887 > 0$$

$$D(q||p) = \frac{1}{2}\log\left(\frac{1/2}{1/4}\right) + \frac{1}{2}\log\left(\frac{1/2}{3/4}\right) = \frac{1}{2}\left(1 - \log\frac{3}{2}\right) \approx 0.2973 > 0$$

Note that $D(p||q) \neq D(q||p)$.

Proof of Property 2 (Information Inequality): To prove this property, we will use the following fact:

Identity:

$$\log_2 y \leq \frac{y-1}{\log_e 2},\ \forall\ y\ \in\ \mathbb{R}$$

Proof of identity: First, note that the following is true:

$$1 + x \leq e^x,\ \forall\ x\ \in\ \mathbb{R} \tag{3}$$

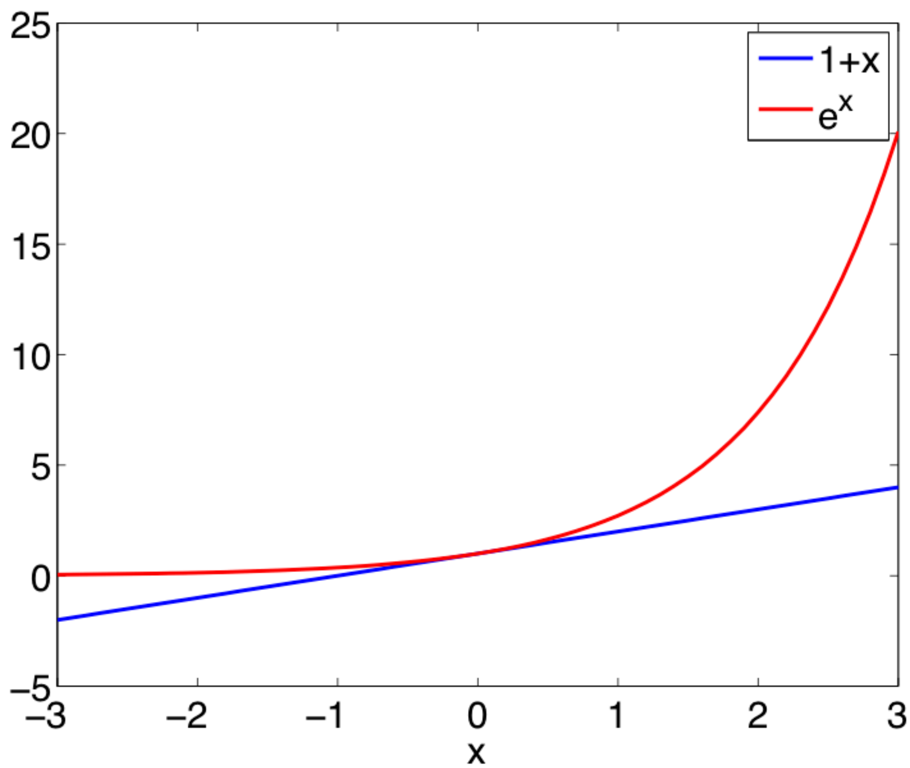

Figure 1: $1 + x$ and e^x

Taking the natural log of both sides of (3),

$$\log_e(1 + x) \leq x$$

Using the fact that $\log_e T = \log_e 2 \log_2 T$ (change of base formula),

$$\begin{aligned}
\log_e 2 \log_2 (1+x) &\leq x \\
\log_2 (1+x) &\leq \frac{x}{\log_e 2}
\end{aligned}$$

Let $y = x + 1$:

$$\log_2 y \leq \frac{y-1}{log_e 2}$$

Going back to the proof of Information Inequality,

$$\begin{aligned}
D(p||q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= -\sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \\
&\geq -\sum_{x \in \mathcal{X}} p(x) \left(\frac{1 - \frac{q(x)}{p(x)}}{\log_e 2} \right) \\
&= -\frac{1}{\log_e 2} \sum_{x \in \mathcal{X}} (p(x) - q(x)) = 0 \\
&\implies D(p||q) \geq 0
\end{aligned}$$

Identity: Let $X \in \mathcal{X}$, $|\mathcal{X}| = M$. Then

$$\log M - H(X) \geq 0$$

Proof: As stated previously, the maximum value of entropy is $\log |\mathcal{X}| = \log M$, which occurs when X is uniformly-distributed. Now we can prove this by KL-distance.

$$\begin{aligned}
\log M - H(X) &= \sum_{x \in \mathcal{X}} p(x) \log M + \sum_{x \in \mathcal{X}} p(x) \log p(x) \\
&= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{1/M} \\
&= D(p|| \tfrac{1}{M}) \geq 0
\end{aligned}$$

Note that the inequality above is known to us by the non-negativity property of KL-distance.

If the distribution of X is uniform, where $p(X = x) = \frac{1}{M}$ $\forall$ $x \in \mathcal{X}$, then $D(p|| \frac{1}{M}) = 0$.

3 Mutual Information

Definition: *Mutual information* is defined by

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&\triangleq D\Big(p(x,y) \mid\mid p(x)p(y)\Big)
\end{aligned}$$

Proof:

$$\begin{aligned}
D\Big(p(x,y)\ ||\ p(x)p(y)\Big) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x)p(y|x)}{p(x)p(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(y|x)}{p(y)} \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= H(Y) - H(Y|X)
\end{aligned}$$

If X and Y are independent, $I(X;Y) = 0$.