CFCS
Entropy and Kullback-Leibler Divergence

Miles Osborne (originally: Frank Keller)

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

February 25, 2008

Kullback-Leibler Divergence

Definition: Kullback-Leibler Divergence

For two probability distributions $f(x)$ and $g(x)$ for a random variable X, the Kullback-Leibler divergence or relative entropy is given as:

$$D(f||g) = \sum_{x \in X} f(x) \log \frac{f(x)}{g(x)}$$

The KL divergence compares the entropy of two distributions over the same random variable.

Intuitively, the KL divergence number of additional bits required when encoding a random variable with a distribution $f(x)$ using the alternative distribution $g(x)$.

Kullback-Leibler Divergence

Theorem: Properties of the Kullback-Leibler Divergence

1. $D(f||g) \geq 0$;
2. $D(f||g) = 0$ iff $f(x) = g(x)$ for all $x \in X$;
3. $D(f||g) \neq D(g||f)$;
4. $I(X;Y) = D(f(x,y)||f(x)f(y))$.

So the mutual information is the KL divergence between $f(x,y)$ and $f(x)f(y)$. It measures how far a distribution is from independence.

Kullback-Leibler Divergence

Example

For a random variable $X = \{0, 1\}$ assume two distributions $f(x)$ and $g(x)$ with $f(0) = 1 - r$, $f(1) = r$ and $g(0) = 1 - s$, $g(1) = s$:

$$\begin{aligned} D(f||g) &= (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s} \\ D(g||f) &= (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r} \end{aligned}$$

If $r = s$ then $D(f||g) = D(g||f) = 0$. If $r = \frac{1}{2}$ and $r = \frac{1}{4}$:

$$\begin{aligned} D(f||g) &= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} = 0.2075 \\ D(g||f) &= \frac{3}{4}\log\frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{2}} = 0.1887 \end{aligned}$$

Entropy and Information

Definition: Entropy

If X is a discrete random variable and $f(x)$ is the value of its probability distribution at x, then the entropy of X is:

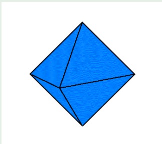
$$H(X) = -\sum_{x \in X} f(x) \log_2 f(x)$$

- Entropy is measured in bits (the log is $\log_2$);
- intuitively, it measures amount of information (or uncertainty) in random variable;
- it can also be interpreted as the average length of message to transmit an outcome of the random variable;
- note that $H(X) \geq 0$ by definition.

Entropy and Information

Example: 8-sided die

Suppose you are reporting the result of rolling a fair eight-sided die. What is the entropy?



The probability distribution is $f(x) = \frac{1}{8}$ for $x = 1 \ldots 8$. Therefore entropy is:

$$\begin{aligned} H(X) &= -\sum_{x=1}^{8} f(x)\log f(x) = -\sum_{x=1}^{8} \frac{1}{8}\log\frac{1}{8} \\ &= -\log\frac{1}{8} = \log 8 = 3 \text{ bits} \end{aligned}$$

This means the average length of a message required to describe (encode) the outcome of the roll of the die is 3 bits.

Entropy and Information

Example: 8-sided die

Suppose you wish to send the result of rolling the die. What is the most efficient way to encode the message?

The entropy of the random variable is 3 bits. That means the outcome of the random variable can be encoded as 3 digit binary message:

1	2	3	4	5	6	7	8
001	010	011	100	101	110	111	000

Example: simplified Polynesian

Example: simplified Polynesian

Polynesian languages are famous for their small alphabets. Assume a language with the following letters and associated probabilities:

x	p	t	k	a	i	u
f(x)	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

What is the per-character entropy for this language?

$$\begin{aligned} H(X) &= -\sum_{x\in\{p,t,k,a,i,u\}} f(x)\log f(x) \\ &= -(4\log\frac{1}{8} + 2\log\frac{1}{4}) = 2\frac{1}{2} \text{ bits} \end{aligned}$$

Example: simplified Polynesian

Example: simplified Polynesian

Now let's design a code that takes $2\frac{1}{2}$ bits to transmit a letter:

p	t	k	a	i	u
100	00	101	01	110	111

Any code is suitable, as long as it uses two digits to encode the high probability letters, and three digits to encode the low probability letters.

Properties of Entropy

Theorem: Entropy

If X is a binary random variable with the distribution $f(0) = p$ and $f(1) = 1 - p$, then:

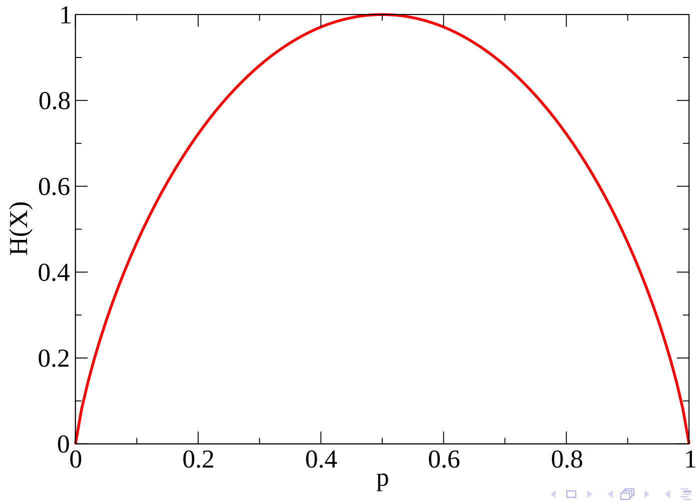
- $H(X) = 0$ if $p = 0$ or $p = 1$
- $\max H(X)$ for $p = \frac{1}{2}$

Intuitively, an entropy of 0 means that the outcome of the random variable is determinate; it contains no information (or uncertainty).

If both outcomes are equally likely (p $= \frac{1}{2}$), then we have maximal uncertainty.

Properties of Entropy

Visualize the content of the previous theorem:



Joint Entropy

Definition: Joint Entropy

If X and Y are discrete random variables and $f(x, y)$ is the value of their joint probability distribution at (x, y), then the joint entropy of X and Y is:

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

The joint entropy represents the amount of information needed on average to specify the value of two discrete random variables.

Conditional Entropy

Definition: Conditional Entropy

If X and Y are discrete random variables and $f(x, y)$ and $f(y|x)$ are the values of their joint and conditional probability distributions, then:

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

is the conditional entropy of Y given X.

The conditional entropy indicates how much extra information you still need to supply on average to communicate Y given that the other party knows X.

Conditional Entropy

Example: simplified Polynesian

Now assume that you have the joint probability of a vowel and a consonant occurring together in the same syllable:

$f(x, y)$	p	t	k	$f(y)$
a	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{1}{16}$	$\frac{1}{2}$
i	$\frac{1}{16}$	$\frac{3}{16}$	0	$\frac{1}{4}$
u	0	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
$f(x)$	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$	

Compute the conditional probabilities; for example:

$$f(a|p) = \frac{f(a, p)}{f(p)} = \frac{\frac{1}{16}}{\frac{1}{8}} = \frac{1}{2}$$

$$f(a|t) = \frac{f(a, t)}{f(t)} = \frac{\frac{3}{8}}{\frac{3}{4}} = \frac{1}{2}$$

Conditional Entropy

Example: simplified Polynesian

Now compute the conditional entropy of a vowel given a consonant:

$$\begin{aligned}
H(V|C) &= -\sum_{x \in C} \sum_{y \in V} f(x, y) \log f(y|x) \\
&= -(f(a, p) \log f(a|p) + f(a, t) \log f(a|t) + f(a, k) \log f(a|k) + \\
&\quad f(i, p) \log f(i|p) + f(i, t) \log f(i|t) + f(i, k) \log f(i|k) + \\
&\quad f(u, p) \log f(u|p) + f(u, t) \log f(u|t) + f(u, k) \log f(u|k)) \\
&= -(\tfrac{1}{16} \log \tfrac{\frac{1}{16}}{\frac{1}{8}} + \tfrac{3}{8} \log \tfrac{\frac{3}{8}}{\frac{3}{4}} + \tfrac{1}{16} \log \tfrac{\frac{1}{16}}{\frac{1}{8}} + \\
&\quad \tfrac{1}{16} \log \tfrac{\frac{1}{16}}{\frac{1}{8}} + \tfrac{3}{16} \log \tfrac{\frac{3}{16}}{\frac{3}{4}} + 0 + \\
&\quad 0 + \tfrac{3}{16} \log \tfrac{\frac{3}{16}}{\frac{3}{4}} + \tfrac{1}{16} \log \tfrac{\frac{1}{16}}{\frac{1}{8}}) \\
&= \tfrac{11}{8} = 1.375 \text{ bits}
\end{aligned}$$

Conditional Entropy

For probability distributions we defined:

$$f(y|x) = \frac{f(x,y)}{g(x)}$$

A similar theorem holds for entropy:

Theorem: Conditional Entropy

If X and Y are discrete random variables with joint entropy $H(X,Y)$ and the marginal entropy of X is $H(X)$, then:

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

Division instead of subtraction as entropy is defined on logarithms.

Conditional Entropy

Example: simplified Polynesian

Use the previous theorem to compute the joint entropy of a consonant and a vowel. First compute $H(C)$:

$$\begin{aligned} H(C) &= -\sum_{x \in C} f(x) \log f(x) \\ &= -(f(p)\log f(p) + f(t)\log f(t) + f(k)\log f(k)) \\ &= -(\frac{1}{8}\log\frac{1}{8} + \frac{3}{4}\log\frac{3}{4} + \frac{1}{8}\log\frac{1}{8}) \\ &= 1.061 \text{ bits} \end{aligned}$$

Then we can compute the joint entropy as:

$$H(V,C) = H(V|C) + H(C) = 1.375 + 1.061 = 2.436 \text{ bits}$$

Summary

- the Kullback-Leibler divergence is the distance between two distributions (the cost of encoding $f(x)$ through $g(x)$).
- Entropy measures the amount of information in a random variable or the length of the message required to transmit the outcome;
- joint entropy is the amount of information in two (or more) random variables;
- conditional entropy is the amount of information in one random variable given we already know the other.