CS 475 Machine Learning: Lecture 6
Information Theory

1 Information Theory

Information theory is the study of the transmission of bits across a noisy channel.
The currency of information theory is "bits"
How many bits do I need to encode information?
The model is a channel with a sender and receiver. I want to send you information. How many bits do I need to do it? How expensive is information?

I have a coin {Heads, Tails}. I want to send you the result of the coin flip. On average, how many bits do I need? (1 bit)

- Heads - $\langle 0 \rangle$
- Tails - $\langle 1 \rangle$

Of course, not everything fits into 1 bit.
Horse race with 4 horses. How many bits? (2 bits)

- Horse A - $\langle 0, 0 \rangle$
- Horse B - $\langle 0, 1 \rangle$
- Horse C - $\langle 1, 0 \rangle$
- Horse D - $\langle 1, 1 \rangle$

Let's say the sender and receiver know extra information.
Distribution over each horse winning the race.

- Horse A - $\frac{1}{2}$
- Horse B - $\frac{1}{4}$
- Horse C - $\frac{1}{8}$
- Horse D - $\frac{1}{8}$

Can we do better than 2 bits?

- Horse A - $\langle 0 \rangle$
- Horse B - $\langle 1, 0 \rangle$
- Horse C - $\langle 1, 1, 0 \rangle$
- Horse D - $\langle 1, 1, 1 \rangle$

Notice that I now have up to 3 bits, but only for unlikely events.
How many on average?
$.5 \times 1 + .25 \times 2 + .125 \times 3 + .125 \times 3 = 1.75$ bits

1.1 Entropy

In information theory, entropy is the uncertainty associated with a random value.
We can ask how uncertain are we with the random value (horse race) we are receiving.
The expected value (number of bits) in the message.

The entropy of a discrete random variable X is:

$$H(X) = E(I(X))$$

E is the expected value function
$I(X)$ is the information content of the message/random variable X
We can write this out as:

$$H(X) = \sum_{i=1}^{n} p(x_i)I(x_i) = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i)$$

First part- weigh each event's information by the probability that it occurs
Second part- the amount of bits needed to store the information.
Consider the horse race. For an event that occurs $\frac{1}{2}$ the time we need:

$$-\log_2 p(x_i) = -\log_2 \frac{1}{2} = 1\text{bit}$$

For an event that occurs $\frac{1}{4}$ the time we need:

$$-\log_2 p(x_i) = -\log_2 \frac{1}{4} = 2\text{bits}$$

So to know how much information we need for the horse race, use the entropy of the message:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = 1.75\text{bits}$$

1.2 Notes on Entropy

High entropy- the distribution is uniform. We can't predict which events will happen. More bits needed.
Low entropy- the distribution is peaked. We can predict which events will happen. Less bits needed.

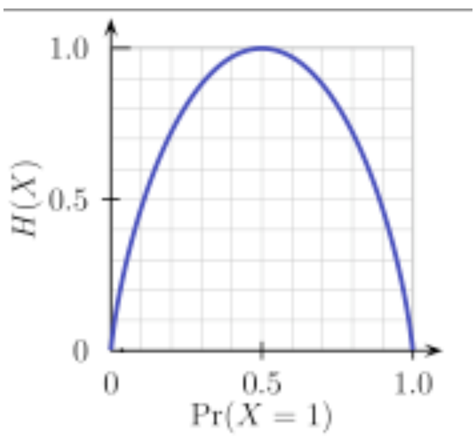

Figure shows the entropy for a coin. If the coin has equal probability of heads vs. tails, then high entropy (full bit needed). Otherwise, less bits.

1.3 Conditional Entropy

What if we both already know some information. How many more bits are needed?
Example, you knew that horse A or B won, but not sure which. Do I still need 1.75 bits? Obviously not.
Define $H(Y|X = x)$- the number of bits needed to send Y given that we both know $X = x$.
Its the same as entropy but for only the cases when $X = x$.
The full expected condition entropy is $H(Y|X)$ where we average over all the values that X can take in $H(Y|X = x)$.

$$\begin{aligned}
H(Y|X) &= \sum_{x \in X} p(x) H(Y|X = x) & (1)\\
&= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x) & (2)\\
&= -\sum_{x \in X} \sum_{y \in Y} p(y, x) \log p(y|x) & (3)\\
&= -\sum_{x \in X, y \in Y} p(y, x) \log p(y|x) & (4)\\
&= -\sum_{x \in X, y \in Y} p(y, x) \log \frac{p(y, x)}{p(x)} & (5)\\
& & (6)
\end{aligned}$$

1.4 Information Gain

Now that we can 1) quantify how much information is in a message and 2) how much that reduces when both sides know information:
We can talk about information savings.
I want to send Y with as few bits as possible. How many bits could I save if we both knew X?
In terms of horse race: I want to say that horse A won the race, how many bits would I save if we both knew it was horse A or B?
Information gain: how much information have we gained if you knew X?

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

Intuitively, X has a high information gain with respect to Y if, knowing X , it takes many fewer bits to transmit Y.