Statistical Methods in Natural Language Processing (NLP)

Class 7: Information Theory Basic Concepts

Charalambos (Haris) Themistocleous

Department of Philosophy, Linguistics and Theory of Science, Centre for Linguistic Theory and Studies in Probability

Introduction

- Introduction to Information Theory.
- Entropy
- Joint Entropy
- Conditional Entropy
- Mutual information
- Noisy Channels

Information Theory

Information Theory was founded by Claude Elwood Shannon (April 30, 1916 February 24, 2001) in his landmark paper, "A Mathematical Theory of Communication" (1948).

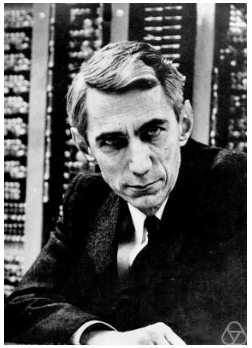

Figure: Claude Elwood Shannon (April 30, 1916 February 24, 2001)

Information Theory

- Maximizing the amount of information that can be transmitted over an imperfect communication channel.
- **Entropy** is the average uncertainty of a random variable.
- Entropy is measured in **bits**.
- More information = less entropy.
- A random variable with only one value: a metal ball that always falls down and never goes up has no uncertainty and its entropy is defined as 0.
- You can say that we do not get information from this.
- A fair coin that can be heads or tails has entropy 1.
- The roll of a fair four-sided dice has 2 bits of entropy, because it takes two bits to describe one of four equally probable choices.

Entropy

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

Entropy Boolean random variable

Let B(q) be the entropy of a Boolean random variable that is true with probability q: $B(q) = -(q log_2 q + (1-q) log_2 (1-q))$

Entropy

1. $p(x)$ is the probability mass function of a random variable X, over a discrete set of symbols X :
2. $p(x) = P(X = x), x \in X$
3. The probability of getting heads when we toss two coins is {tt,hh,th,ht}: p(0) = 1/4, p(1)= 1/2, p(2) = 1/4.
4. X is a discrete random variable, $p(X)$
5. log_2 so in this analysis 0 log 0 = 0.

Examples

What is the entropy of a fair coin:

- $H(Fair) = (0.5 log_2 0.5 + 0.5 log_2 0.5) = 1$
- in R: (-(log2(0.5)*.5+log2(0.5)*.5))
- If the coin is modified to give 79% heads, then:
- $H(notFair) = (0.79 log_2 0.79 + 0.01 log_2 0.01) \approx 0.74 bits$
- In R: -(log2(0.79)*.79+log2(0.21)*.21)

Example

What is the entropy of an 8 sided die?

Example

What is the entropy of an 8 sided die?

$$H(p) = H(X) = -\sum_{x \in X} p(x) log_2 p(x) = -\sum_{i=8}^{8} \frac{1}{8} log_2 \frac{1}{8} = log\ 8 = 3 bits.$$

Note the summation part becomes 1.

Consequences

If an the information is 3 bits it means that the whole information can be sent using 3 digit binary messages:

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

Example

Example from Manning and Schutze (2001:62). Simplified Polynesian appears to be just a random sequence of letters, with the following letters frequencies:

p	t	k	a	i	u
1/8	1/4	1/8	1/4	1/8	1/8

Example

$$H(P) = -\sum_{i\in\{p,t,k,a,i,u\}} P(i)log_2P(i)$$
$$= -[4 \times \frac{1}{8}log_2\frac{1}{8} + 2 \times \frac{1}{4}log\frac{1}{4}]$$
$$= 2\frac{1}{2}bits.$$

So we can design a code that can transmit a letter that takes 2 1/2 bits.

- if the code starts with 0 is length 2
- if the code starts with 1 is length 3

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

Joint Entropy

The joint entropy of 2 response variables $X, Y \sim p(x, y)$ is the amount of the information needed on average to specify both their values and it is calculated in the following way:

$$H(X, Y) = -\sum_{x\in X}\sum_{y\in Y} p(x, y)log_2p(x, y)$$

Conditional Entropy

The conditional entropy of 2 response variables $X, Y \sim p(x, y)$ is the amount of the information needed on average to specify both their values and it is calculated in the following way:

Conditional Entropy

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

Mutual Information

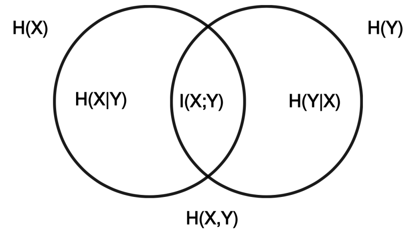

Figure: Relationship between mutual information and entropy.

Surprise

1. When a model captures more of the structure of a language then the entropy should be lower than a model that captures less structure of a language.
2. Pointwise entropy as a measure of surprise: $H(w|h) = -log_2 m(w|h)$ where w is a next word, h is the what we already know and m is the model of the distribution of a certain language.
3. If two words appear usually next to each other, e.g., Costa Rica, then the amount of surprise is very small or close to zero $(-log_2 = 0)$ but if the model estimates that two words w cannot follow: as in ***cat the** then $m(w|h) = 0$.

The overall surprise in a model is the sum of all the words:

$$
\begin{aligned}
H_{total} &= \sum_{j=1}^{n} log_2 m(w_j | w_1, w_2, \ldots, w_{j-1}) \\
&= -log_2 m(w_1, w_2, \ldots, w_n)
\end{aligned}
$$

Applications

1. Ngram models or Markov chains
2. Machine Learning

Next Class

- Machine Learning
- Introduction to basic algorithms
- Training and test sets
- Model Evaluation