Information Theory lecture 3

COMSM0075 Information Processing and Brain

comsm0075.github.io

September 2020

Joint and conditional entropy

Typically we want to use information theory to study the relationship between two random variables.

Joint entropy

Given two random variables X and Y the probability of getting the pair (x_i, y_j) is given by the **joint probability** $p_{(X,Y)}(x_i, y_j)$. The **joint entropy** is just the entropy of the joint distribution:

$$H(X,Y) = -\sum_{i,j} p_{X,Y}(x_i, y_j) \log_2 p_{X,Y}(x_i, y_j)$$

Joint entropy

Given two random variables X and Y the probability of getting the pair (x_i, y_j) is given by the **joint probability** $p_{(X,Y)}(x_i, y_j)$. The **joint entropy** is just the entropy of the joint distribution:

$$H(X,Y) = -\sum_{i.j} p_{X,Y}(x_i, y_j) \log_2 p_{X,Y}(x_i, y_j)$$

An example

	x_0	x_1
y_0	$1/4$	$1/4$
y_1	$1/2$	0

The joint entropy

	x_0	x_1
y_0	$1/4$	$1/4$
y_1	$1/2$	0

$$H(X,Y) = -\frac{1}{2}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2} = \frac{3}{2}$$

Conditional probability

$p_{X|Y}(x_i|y_j)$ is the **conditional probability** of x_i given y_j; if we know $Y = y_j$ it gives the probability that the pair is (x_i, y_j).

Conditional probability

$$p_{(X,Y)}(x_i, y_j) = p_{X|Y}(x_i|y_j)p_Y(y_j)$$

Conditional probability

$$p_{(X,Y)}(x_i, y_j) = p_{X|Y}(x_i|y_j)p_Y(y_j)$$

Conditional probability

$$p_{(X,Y)}(x_i, y_j) = p_{X|Y}(x_i|y_j)p_Y(y_j)$$

Conditional probability

$$p_{(X,Y)}(x_i, y_j) = p_{X|Y}(x_i|y_j)p_Y(y_j)$$

Conditional probability

$$p_{X|Y}(x_i|y_j) = \frac{p_{(X,Y)}(x_i, y_j)}{p_Y(y_j)}$$

Marginal probabilities

$$p_X(x_i) = \sum_j p_{(X,Y)}(x_i, y_j)$$

The conditioned entropy

So lets substitute the conditional probability into the formula for the entropy

$$H(X|Y = y_j) = -\sum_i p_{X|Y}(x_i|y_j) \log_2 p_{X|Y}(x_i|y_j)$$

This is the entropy of X is we know $Y = y_j$; we'll call this the **conditioned entropy**.

This can go either way!

The previous example:

	x_0	x_1
y_0	1/4	1/4
y_1	1/2	0

has conditional distributions for $Y = y_0$:

	x_0	x_1
$Y = y_0$	1/2	1/2

and for $Y = y_1$:

	x_0	x_1
$Y = y_1$	1	0

This can go either way!

	x_0	x_1
$Y = y_0$	$1/2$	$1/2$

so

$$H(X|Y = y_0) = 1$$

	x_0	x_1
$Y = y_1$	1	0

so

$$H(X|Y = y_1) = 0$$

The conditional entropy

The **conditional entropy** is the average conditioned entropy:

$$H(X|Y) = \sum_j p_Y(y_j) H(X|Y = y_j)$$

The conditional entropy

The **conditional entropy** is the average conditioned entropy:

$$H(X|Y) = \sum_j p_Y(y_j) H(X|Y = y_j)$$

It tells us how much information there is in X *on average* if you know Y, averaged over the possible outcomes of 'knowing Y'

The conditional entropy

The **conditional entropy** is the average conditioned entropy:

$$H(X|Y) = \sum_j p_Y(y_j) H(X|Y = y_j)$$

so substituting in for $H(X|Y = y_j)$

$$H(X|Y) = -\sum_{i,j} p_Y(y_j) p_{X|Y}(x_i|y_j) \log_2 p_{X|Y}(x_i|y_j)$$

and, since $p_Y(y_j) p_{X|Y}(x_i, y_j) = p_{(X,Y)}(x_i, y_j)$, we have

$$H(X|Y) = -\sum_{i,j} p_{X,Y}(x_i, y_j) \log_2 p_{X|Y}(x_i|y_j)$$

The conditional entropy

$H(X|Y)$ *is the average amount of information still in* X *when we know* Y.

The conditional entropy has nice properties

If X and Y are independent then

$$p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_j)$$

for all i and j and

$$p_{X|Y}(x_i|y_j) = p_X(x_i)$$

so

$$H(X|Y) = -\sum_{i,j} p_{X,Y}(x_i, y_j) \log_2 p_{X|Y}(x_i|y_j) = H(X)$$

The conditional entropy has nice properties

Conversely, if X is determined by Y, for example if the only (x_j, y_i) pairs that actually occur are (x_i, y_i). In this case $p_{X|Y}(x_j|y_i)$ is zero for every x_i except $p_{X|Y}(x_i|y_i) = 1$. In this case

$$H(X|Y) = 0$$

Conditional entropy example

	x_0	x_1
y_0	$1/4$	$1/4$
y_1	$1/2$	0

with $H(X|Y = y_0) = 1$ and $H(X|Y = y_1) = 0$.

Conditional entropy example

	x_0	x_1
y_0	1/4	1/4
y_1	1/2	0

with $H(X|Y=y_0)=1$ and $H(X|Y=y_1)=0$. The marginal distribution $p_Y(y)$ is

	y_0	y_1
$p_Y(y)$	1/2	1/2

and hence

$$H(X|Y)=\frac{1}{2}\times 1+\frac{1}{2}\times 0=\frac{1}{2}$$

Conditional entropy example

	x_0	x_1
y_0	$1/4$	$1/4$
y_1	$1/2$	0

The other marginal distribution $p_X(x)$ is

	x_0	x_1
$p_X(x)$	$3/4$	$1/4$

and hence

$$H(X) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.81$$

Conditional entropy example

Hence

$$H(X|Y) < H(X)$$

Conditional entropy is less than the entropy

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

which is as it should be!

A chain rule

This is what you get from the definition of entropy if you use

$$p_{X,Y}(x_i, y_j) = p_{X|Y}(x_i|y_j)p_Y y_j$$

So take

$$H(X, Y) = -\sum_{i,j} p_{X,Y}(x_i, y_j) \log_2 p_{X,Y}(x_i, y_j)$$

and substitute for the $p_{X,Y}(x_i, y_j)$ inside the log. A bit of mathematics gives you

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

A chain rule

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

This again makes sense; the amount of information in X and Y is the amount of information in X plus the amount of information remaining in Y if we already know X.