Class no 7:

Joint entropy in 2 RVs

$$H(X,Y) = \sum_{(x,y) \in supp(P_{X,Y})} p(x,y) \log_2 \frac{1}{p(x,y)}$$

$\mathcal{X} \times \mathcal{Y} \triangleq \{(x,y) : x \in \mathcal{X}, y \in \mathcal{Y}\}$

"$\times$" indicates Cartesian product

$P_{X,Y} : \boxed{\mathcal{X} \times \mathcal{Y}} \to [0,1]$

Joint distribution of X & Y

$$\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x,y) = 1$$

Note:

Suppose $X = Y$. Then $H(X,Y) = H(X) = H(Y)$

If X, Y independent $H(X,Y) = H(X) + H(Y)$

Joint entropy in RVs $X_1, \ldots, X_n$

$$H(X_1, \ldots, X_n) \triangleq \sum_{(x_1, \ldots, x_n) \in supp(P_{X_1, \ldots, X_n})} p_{X_1 \cdots X_n}(x_1, \ldots, x_n) \log_2 \frac{1}{p(x_1, \ldots, x_n)}$$

Chain Rule for Joint Entropy:

Lemma:

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1,X_2) + \ldots + H(X_n|X_1, \ldots, X_{n-1})$$

entropy of X_3 given (X_1 & X_2)

Where

$H(X,Y|V,U)$ Conditional Joint entropy of X & Y given U & V

(see next page)

Recall:

$$H(X|Y) = \sum_{y \in supp(P_Y)} P_Y(y) \underbrace{H(X|Y=y)}$$

entropy of X wrt distribution $P_{X|Y}(x|y)$

$p(X=x|Y=y)$

$$H(X|Y=y) = \sum_{x \in supp(P_{X|Y=y})} p(x|y) \log \frac{1}{p(x|y)}$$

$$H(\underline{X,Y} \mid \underline{U,V}) \triangleq \sum_{\substack{(u,v) \in supp(P_{U,V})}} P_{U,V}(u,v)\, H(X,Y \mid U=u, V=v)$$

(u,v) ↓ small (u,v)

where

$$H(X,Y \mid U=u, V=v) \triangleq \sum_{(x,y) \in supp(P_{X,Y \mid U=u, V=v})} p(x,y \mid u,v) \log_2 \frac{1}{p(x,y \mid u,v)}$$

Note that $p(x,y \mid u,v)$ is a valid joint distribution on RVs X & Y.

This can be extended to any number of variables before & after the conditioning.

Note:

How to understand $p(x_1, x_2, x_3 \mid y_1, y_2)$?

Recall:

$$p_{X_1,X_2}(x_1, x_2) = p_{X_1}(x_1)\, p_{X_2, X_1}(x_2 \mid x_1)$$

$$\left[P_{Y|X}(y|x) \triangleq \frac{p(x,y)}{p(x)} \right]$$

$$p(x_1, x_2, x_3 \mid y_1, y_2) = p(x_1, x_2 \mid y_1, y_2)\, p(x_3 \mid x_1, x_2, y_1, y_2)$$

$$= \frac{p(x_3, y_2 \mid y_1)\, p(x_1, x_2 \mid x_3, y_1, y_2)}{p(y_2 \mid y_1)}$$

$$= p(x_3, y_2, x_1, x_2 \mid \underline{y_1}) / p(y_2 \mid \underline{y_1})$$

$$= p(x_1, x_2, x_3 \mid y_2, y_1)$$

Now we prove the chain rule:

Proof: Note the chain rule wrt the joint distribution $p(x_1, \ldots, x_n)$

$$= p(x_1)\, p(x_2, \ldots, x_n \mid x_1)$$

$$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3, \ldots, x_n \mid x_1, x_2)$$

$$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2)\, p(x_4, \ldots, x_n \mid x_1, x_2, x_3)$$

$$= \cdots$$

$$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2) \cdots p(x_n \mid x_1, \ldots, x_{n-1})$$

Use defn of joint entropy $H(X_1 \ldots X_n)$ to complete the proof [Exercise]

Class no 8:

$X \longrightarrow$ entropy of $H(X) \Rightarrow$ avg uncertainty abt X

Observer (R_X) sees 'X'

$\rightarrow$ How much uncertainty does Obs (R_X) have abt X?

Uncertainty abt X after Obs $X = H(X \mid X)$
$= 0$

$\rightarrow$ What is the original uncert in X (before observing X) $= H(X)$
↳ (assume that P_X is known)

$\Rightarrow$ Reduction in avg uncertainty of X achieved by observing X

$$= \underset{\text{original}}{H(X)} - \underset{\text{uncert after obs}}{H(X|X)} = H(X)$$

Suppose X, Y are related to each other

(captured by a given joint prob. distribution $P_{X,Y}(x,y) = P(X=x, Y=y), \forall x, y$)

(obs know this)

Reduction in uncert. of X after observing Y

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

avg uncertainty in X (left in X) after obs. Y.

"Information gained abt X after observing Y"

"MUTUAL INFORMATION BETWEEN X & Y", $I(X;Y)$

$$I(X;Y) \triangleq H(X) - H(X|Y).$$

Easy to show that $H(X) - H(X|Y) = H(Y) - H(Y|X)$

also equivalent to Info gained abt Y after observing X.

(1) $I(X;Y) \leq \min(H(X), H(Y))$

is important because it tells that the information gained depends on 'grosser' ~~'quantized'~~ observation.

(2) $I(X;Y) \geq 0$ (Exercise: Represent $I(X;Y)$ as a relative entropy between two prob distributions)

— x —

Information theory

Efficient Source Representation (Some Random variable) (or Seq of RVs)

(Data Compression)

[Source Coding]

High Rate & High fidelity (low prob of error) communication via some Noisy channel

[Channel Coding]

Overview of Source Coding:

Suppose we have $X \in \{a, b\}$ a binary source with prob distribution P_X

Suppose Obs observes one instance of X, then wants to store (or communicate) it through some noise-free medium, which can carry/store only $\{0,1\}$ (bits)

In general we need 1 bit. $a \to 0$, $b \to 1$ (example)

But if Rx knows P_X, & it happens that

$$P_X(b) = 1, \quad P_X(a) = 0,$$

$P_X(b) \to (P(X=b))$, $P_X(a) \to P(X=a)$

then Rx need not even 'Read' or 'Receive' the encoded value to 'know' X. It can simply declare the value of X to be b & this will be correct/error-free with probability = 1.

$\Rightarrow$ Our code is not Required at all

$\Rightarrow$ we need a 0-length code.

Now, Suppose we are allowing for a small probability of error $\Rightarrow p(\text{error}) \leq \epsilon$ for small $\epsilon \in [0,1)$

Note that if length of code (stored symbols) $= 1$, then $P(\text{error}) = 0$.

Can we have length $= 0$, & some prob distribution for X for which $p(\text{error}) \leq \epsilon$?

Ans: Yes!

Suppose $P(X=b) = 1-\epsilon$, $P(X=a) = \epsilon$.

For this suppose we use 0-length code,

what is $P(\text{error}) = ?$

$\rightarrow$ "Declare $X=b$ always" at decoder,

this gives us $p(\text{error}) \leq \epsilon$.