Week 3

0.1 Joint Entropy in multiple RV's

$$H(X_1, \cdots, X_n) = \sum_{(x_1,\cdots,x_n)\in supp(P_{X_1,\cdots,X_n})} P(x_1, \cdots, x_n) \log_2 \frac{1}{P(x_1, \cdots, x_n)}$$

Where $P_{X,Y} := \mathcal{X} \times \mathcal{Y} \to$ Cartesian Product.
$\mathcal{X} \times \mathcal{Y} := \{(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}\}$

0.2 Conditional Joint Entropy

$$H(X, Y/U, V) := \sum_{(u,v)\in supp(P_{U,V})} P_{U,V}(u, v) H(X, Y/U = u, V = v)$$

where,

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

This can be extended to any number of variables before and after the conditioning.
eg: $P(x_1, x_2, x_3/y_1, y_2) = P(x_1, x_2/x_3, y_1, y_2) + P(x_3/x_1, x_2, y_1, y_2)$.

1 Chain Rule for Joint Entropy

Lemma:

$$\begin{aligned} H(X_1, \cdots, X_n) = H(X_1) + H(X_2/X_1) + H(X_3/X_1, X_2) + \cdots \\ + H(X_n/X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i/X_{i-1}, \cdots, X_1) \end{aligned} \quad (1)$$

Proof:

$$\begin{aligned} P(x_1, \cdots, x_n) &= P(x_1)P(x_2, \cdots, x_n/x_1) \\ &= P(x_1)P(x_2/x_1)P(x_3, \cdots, x_n/x_1, x_2) \\ &= P(x_1)P(x_2/x_1)P(x_3/x_1, x_2)P(x_4, \cdots, x_n/x_1, x_2, x_3) \\ &= P(x_1)P(x_2/x_1)P(x_3/x_1, x_2) \cdots P(x_n/x_1, \cdots, x_{n-1}) \\ &= \prod_{i=1}^{n} P(x_i/x_{i-1}, \cdots, x_1) \end{aligned}$$

We can substitute this value in the definition for joint entropy and expand, giving us the final result.

$$\begin{aligned}
H(X_1, \cdots, X_n) &= -\sum_{x_1, \cdots, x_n} P(x_1, \cdots, x_n) \log \prod_{i=1}^{n} P(x_i/x_{i-1}, \cdots, x_1) \\
&= -\sum_{x_1, \cdots, x_n} \sum_{i=1}^{n} P(x_1, \cdots, x_n) \log P(x_i/x_{i-1}, \cdots, x_1) \\
&= -\sum_{i=1}^{n} \sum_{x_1, \cdots, x_n} P(x_1, \cdots, x_n) \log P(x_i/x_{i-1}, \cdots, x_1) \\
&= \sum_{i=1}^{n} H(X_i/X_{i-1}, \cdots, X_1)
\end{aligned}$$

2 Mutual Information

We know that the average uncertainity in X is known as entropy of X, and it is given as $H(X)$. If an observer(RX) sees X, the uncertainity in X would become 0 i.e $H(X/X)$.

Hence the reduction in average uncertainity of X achieved by observing X is $H(X) - H(X/X) = H(X)$.

Suppose X, Y are related to each other and the observer knows it's joint probability distribution $P(x, y) = P(X = x, Y = y) \; \forall \; x, y$.

The "reduction in uncertainity of X after observing Y", "information gained about X after observing Y" or "mutual information between X & Y" is:

$$I(X; Y) := H(X) - H(X/Y)$$

By symmetry it follows that:

$$I(X; Y) = H(Y) - H(Y/X) = H(X) + H(Y) - H(X, Y)$$

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

   Information gained depends on the quantized observation.

2. $I(X; Y) \geq 0$

Proof:

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X/Y) \\
&= -\sum_{x \in supp(P_X)} P(x)\log(P(x) - \left(-\sum_{x,y \in suppP_{X,Y}} P(x,y)\log(x/y)\right) \\
&= -\sum_{x,y} P(x,y)\log P(x) + \sum_{x,y} P(x,y)\log P(x/y) \\
&= \sum_{x,y} P(x,y)\log \frac{P(x/y)}{P(x)} \\
&= \sum_{x}\sum_{y} P(x,y)\log \frac{P(x,y)}{P(x)P(y)} \\
&= D(P(x,y)||P(x)P(y))
\end{aligned}
$$

As we know that the relative entropy of 2 probability distributions is positive, the proof is complete.

3 Information Theory

- Efficient source representation.(using some random variable or sequence of R.V's)(data compression and source coding)
- High rate and high fidelity.(low probability of error)(channel coding)

3.1 Source Coding

Suppose we have $X \in \{a, b\}$, a binary source with probability distribution P_X. And an observer observes one instance of X, then wants to store(or communicate) it through a noise-free medium, which can carry or store only $\{0, 1\}$(bits).

But the reciever already knows P_X and it so happens that:

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

Then RX doesn't need to read/recieve the encoded value to know X. It can simply declare the value of X to be b and it will be correct/error-free with probability 1.

We need a 0-length code as code is not required at all.

Suppose we allow a small probability of error $\Rightarrow P(\text{error}) \leq \epsilon, \quad \epsilon \in [0, 1)$.

Then we could have length $= 0$ & $P(\text{error}) \leq \epsilon$, if we let

$P_X(b) = 1 - \epsilon, P_X(a) = \epsilon$.