LECTURE 27

Introduction to information theory

Lecture outline

- Entropy: definitions and properties
- Mutual information: definitions and properties
- How do we achieve entropy mutual information?
- Where to now?

Entropy

- Entropy is a measure of the average uncertainty associated with a random variable
- The entropy of a discrete r.v. X is $H(X) = -\sum_{x\in\mathcal{X}} P_X(x) log_2 \left(P_X(x)\right)$
- entropy is always non-negative
- Joint entropy: the entropy of two discrete r.v.s X, Y with joint PMF $P_{X,Y}(x, y)$ is:

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y) log_2 \left(P_{X,Y}(x,y)\right)$$

- Conditional entropy: expected value of entropies calculated according to conditional distributions $H(Y|X) = E_Z[H(Y|X = Z)]$ for r.v. Z independent of X and identically distributed with X. Intuitively, this is the average of the entropy of Y given X over all possible values of X.

Conditional entropy: chain rule

$$\begin{aligned}
& H(Y|X) = E_Z[H(Y|X = Z)] \\
= & -\sum_{x\in\mathcal{X}} P_X(x) \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) log_2[P_{Y|X}(y|x)] \\
= & -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y) \log_2[P_{Y|X}(y|x)]
\end{aligned}$$

Compare with joint entropy:

$$\begin{aligned}
& H(X,Y) \\
= & -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y) \log_2[P_{X,Y}(x,y)] \\
= & -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y) \log_2[P_{Y|X}(y|x)P_X(x)] \\
= & -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y) \log_2[P_{Y|X}(y|x)] \\
& -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y) \log_2[P_X(x)] \\
= & H(Y|X) + H(X)
\end{aligned}$$

This is the **Chain Rule** for entropy:

$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1 \ldots X_{i-1})$. Question: $H(Y|X) = H(X|Y)$?

Mutual information

Mutual Information: let X, Y be r.v.s with joint PMF $P_{X,Y}(x, y)$ and marginal PMFs $P_X(x)$ and $P_Y(y)$

Definition:

$$\begin{aligned}
& I(X;Y) \\
= & \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y) \log\left(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\right)
\end{aligned}$$

intuitively: measure of how dependent the r.v.s are

Useful expression for mutual information:

$$\begin{aligned}
& I(X;Y) = H(X) + H(Y) - H(X,Y) \\
= & H(Y) - H(Y|X) \\
= & H(X) - H(X|Y) \\
= & I(Y;X)
\end{aligned}$$

Question: what is $I(X;X)$?

How do we achieve entropy and mutual information?

Source coding (compression): we can compress a probabilistic source to rid ourselves of redundancy

Error-free encoding: variable length with average length no better than $H(X) + 1$, but no worse than $H(X)$

Almost error-free encoding: for the typical set, fixed length of $H(X)$ for long enough codewords

How do we achieve entropy and mutual information?

Channel coding theorem (Shannon 1948): we can achieve a rate of information transmission that is arbitrarily close (error-free) to the average mutual information between the input and output of channel

Essence of the proof: the WLLN!

With more patience, we can show strong coding theorem:

$P(\text{message error}) \leq \alpha e^{n(\text{Capacity } - \text{Rate})}$

Converse: we can do no better!

Where to now?

Header course for signal processing, communications, control: 6.011 (you want it all and you want it now)

Communications: 6.450 (you want to see how it's really done)

Data networks: 6.263 (Markov buffs, this is for you!)

Discrete stochastic processes: 6.262 (you like Markov and the SLLN, too)

Detection and estimation: 6.432 (you keep a pet Gaussian pdf in jar under your bed)

Transmission of information (information theory): 6.441 (you use logarithms as stocking stuffers for your friends and family)