LECTURE 1 : Introduction

Handout

Lecture outline

- Course outline, goal and mechanics
- Review of probability
- Define Entropy and Mutual Information
- Questionnaire

Reading: Ch. 1, Scts. 2.1-2.5.

Goals

Our goals in this class are to establish an understanding of the intrinsic properties of transmission of information and the relation between coding and the fundamental limits of information transmission in the context of communications

Our class is not a comprehensive introduction to the field of information theory and will not touch in a significant manner on such important topics as data compression and complexity, which belong in a source-coding class

Probability Space

$(\Omega, \mathcal{F}, P)$

- Ω: Sample space, each $\omega \in \Omega$ is the outcome of an random experiment.
- $\mathcal{F}$: set of events, $E \in \mathcal{F}$, $E \subset \Omega$.
- P: probability measure, $P : \mathcal{F} \to [0, 1]$

Axioms

- $\Omega, \phi \in \mathcal{F}$,
- if $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$,
- if $E_1, E_2, \ldots, \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{F}$.

- $P(\Omega) = 1$,
- $P(E^c) = 1 - P(E)$,
- If $E_1, E_2, \ldots,$ are disjoint, i.e., $E_i \cap E_j = \phi, \forall i \neq j$,

$$P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$$

Example: Flip two coins

$$\begin{aligned} \Omega &= \{\text{HH, HT, TH, TT}\} \\ \mathcal{F} &= \text{ All subsets of } \Omega \end{aligned}$$

$$P(\{\omega\}) = \frac{1}{4}$$

Example: Flip coins until see a head

$$\Omega = \{H, TH, TTH, \ldots,\}$$
$$\mathcal{F} = \text{All subsets of } \Omega$$

$$P(\{\underbrace{TT...T}_{k}H\}) = 2^{k+1}$$

- Mapping between spaces and induced probability measure
- Sometimes we cannot assign probability to each single outcome, e.g., infinite sequence of coin tosses, $P(\{\omega\}) = 0$.

Why we want such probability spaces

Reason 1 :conditional probability

Conditioning on an event $B \in \mathcal{F}$ = change of sample space:

$$\begin{aligned} \Omega &\rightarrow B \\ \text{event } A \in \mathcal{F} &\rightarrow A \cap B \\ P(A) &\rightarrow P(A|B) \end{aligned}$$

Bayes rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- Total probability theorem
- Independent events

Reason 2: random variables

Random Variables

A random variable X is a map from the sample space Ω to $\mathcal{R}$.

The map is deterministic, while the randomness comes only from the underlying probability space.

The map gives a change of probability space and an induced probability measure.

Example $X =$ the number of coin tosses before the first head is seen.

$$P(X = k) = 2^{k+1}, k = 0, 1, \ldots$$

A few things come free:

- Random vectors, $\vec{X}$.
- Function of a random variable, $Y = f(X)$.

Distributions

For a discrete random variable (r.v.),

- Probability Mass Function (PMF),

$$P_X(x) = P(X = x)$$

For a continuous random variable

- Cumulative Distribution Function (CDF),

$$F_X(x) = P(X \leq x)$$

.

- Probability Density Function (PDF),

$$p_X(x) = \frac{d}{dx} F_X(x)$$

expectation, variance, etc.

Entropy

Entropy is

- a measure of the average uncertainty associated with a random variable
- the randomness of a random variable
- the amount of information one obtained by observing the realization of a random variable

Focus on a discrete random variable with n possible values:

$$P(X = x_i) = p_i, i = 1, \ldots, n$$

- The partitioning of the sample space doesn't matter.
- The possible values, x_i, doesn't matter.

Entropy is a function $H(X) = H(p_1, \ldots, p_n)$.

Define Entropy $H(X) = H(p_1, \ldots, p_n)$

Requirement

- H is continuous in $\vec{p}$.
- if $p_i = \frac{1}{n}$, then entropy monotonically increases with n.
- can break into successive choices

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{2}{3}, \frac{1}{3}\right)$$

The only H satisfying these requirements:

$$H = -k \sum_{i=1}^{n} p_i \log_2 p_i$$

Definition

$$
\begin{aligned}
H(X) &= -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x) \\
&= E_P[-\log P(X)]
\end{aligned}
$$

- Entropy is always non-negative.

Example: binary r.v.

$$
X = \begin{cases} 0 & \text{with probability } p \\ 1 & \text{with probability } 1-p \end{cases}
$$

$$
H(X) = -p \log p - (1-p) \log(1-p)
$$

Joint Entropy

The **joint entropy** of two discrete r.v.s X, Y,

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

Example

P	$X=0$	$X=1$
$Y=0$	1/2	1/3
$Y=1$	0	1/6

Example Independent r.v.'s X, Y,

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\,\log_2\Big(P_{X,Y}(x,y)\Big) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\,\log_2\left(P_X(x)P_Y(y)\right) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\,\log_2\left(P_X(x)\right) \\ &\quad -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\,\log_2\left(P_Y(y)\right) \\ &= H(X)+H(Y) \end{aligned}$$

Conditional entropy

Conditional entropy: $H(Y|X)$, a measure of the **average** uncertainty in Y when the realization of X is observed.

- condition on a particular observation $X = x$, $P_Y(y) \rightarrow P_{Y|X}(y|x) = P(Y = y|X = x)$,

  $$H(Y|X = x) = -\sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log P_{Y|X}(y|x)$$

- Average over all possible values of x:

  $$\begin{aligned} H(Y|X) &= -\sum_{x \in \mathcal{X}} P_X(x) H(Y|X = x) \\ &= -\sum_{x,y} P_{XY}(x, y) \log P_{Y|X}(y|x) \\ &= E_{p(X,Y)}[-\log P(Y|X)] \end{aligned}$$

Chain Rule of Entropy

Theorem: Chain Rule

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

Proof

$$\begin{aligned}
& H(X,Y) \\
= & -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\,\mathsf{log}_2[P_{X,Y}(x,y)] \\
= & -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\,\mathsf{log}_2[P_{Y|X}(y|x)P_X(x)] \\
= & -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\,\mathsf{log}_2[P_{Y|X}(y|x)] \\
& -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{X,Y}(x,y)\,\mathsf{log}_2[P_X(x)] \\
= & H(Y|X) + H(X)
\end{aligned}$$

or equivalently,

$$\begin{aligned}
H(X,Y) &= E_{P(X,Y)}[-\mathsf{log}\,P(X,Y)] \\
&= E_{P(X,Y)}[-\mathsf{log}\,P(X) - \mathsf{log}\,P(Y|X)] \\
&= H(X) + H(Y|X)
\end{aligned}$$

By induction

$$H(X_1,\dots,X_n)=\sum_{i=1}^{n} H(X_i|X_1\dots X_{i-1})$$

.

Corollary

X,Y independent $\Rightarrow H(X|Y)=H(X)$

Back to the example

P	$X=0$	$X=1$
$Y=0$	1/2	1/3
$Y=1$	0	1/6

$$H(X,Y)=H\left(\frac{1}{2},\frac{1}{3},\frac{1}{6}\right)$$

notice $0\log 0=0$.

$$\begin{aligned} H(X) &= H\left(\frac{1}{2},\frac{1}{2}\right) \\ H(Y|X) &= \frac{1}{2}H(1,0)+\frac{1}{2}H\left(\frac{2}{3},\frac{1}{3}\right) \end{aligned}$$

Question: $H(Y|X) = H(X|Y)$?

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

or equivalently

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

Definition: Mutual Information

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

The **average** amount of knowledge about X that one obtains by observing the value of Y.

Chain Rule for Mutual Information

Definition: Conditional Mutual Information

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

Chain Rule:

$$\begin{aligned} & I(X_1, X_2; Y) \\ = \; & H(X_1, X_2) - H(X_1, X_2|Y) \\ = \; & H(X_1) + H(X_2|X_1) - H(X_1|Y) - H(X_2|Y, X_1) \\ = \; & I(X_1; Y) + I(X_2; Y|X_1) \end{aligned}$$

By induction

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1 \ldots X_{i-1}, Y)$$