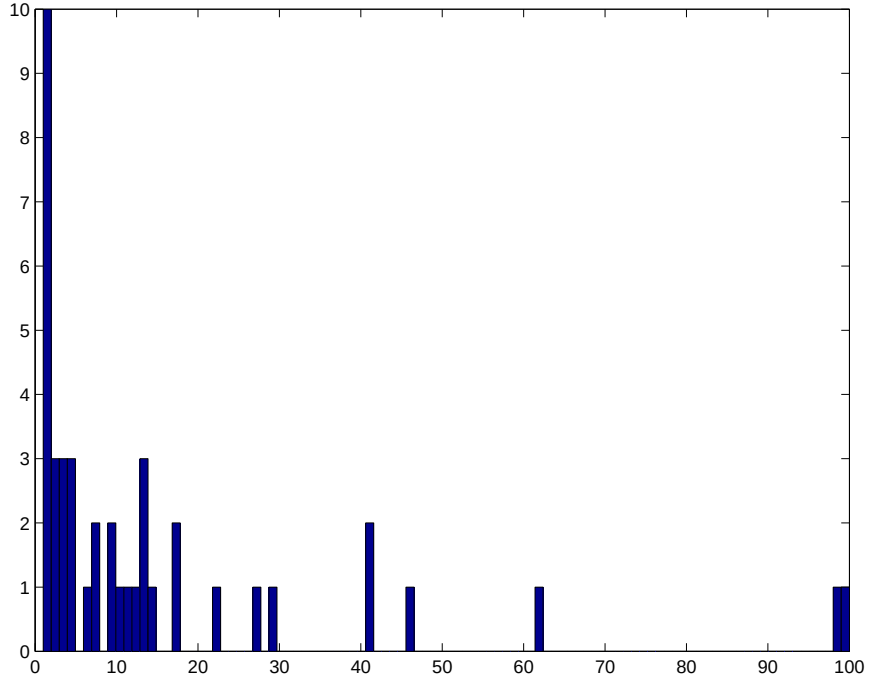
Lecture 2: Entropy and Mutual Information

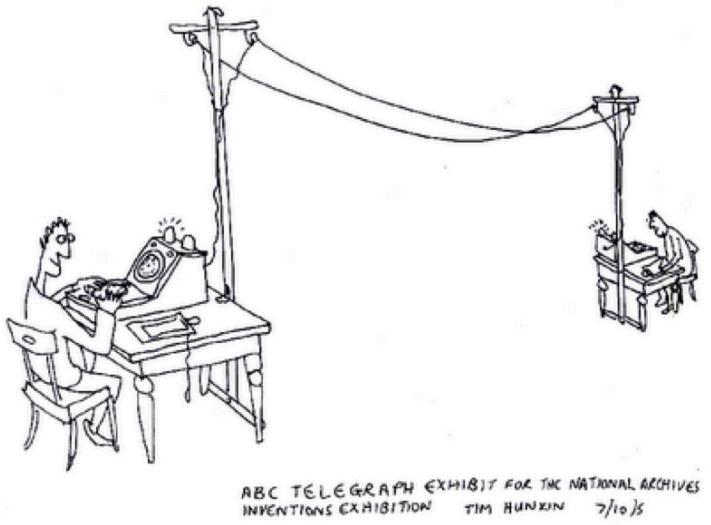
- Entropy
- Mutual Information

The winner is:

Eunsu Ryu, with number 6



A strategy to win the game?



Which horse won?

Uncertainty measure

- Let X be a random variable taking on a finite number M of different values $x_1, \cdots, x_M$
- What is X: English letter in a file, last digit of Dow-Jones index, result of coin tossing, password
- With probability $p_1, \cdots, p_M$, $p_i > 0$, $\sum_{i=1}^{M} p_i = 1$
- Question: what is the uncertainty associated with X?
- Intuitively: a few properties that an uncertainty measure should satisfy
- It should not depend on the way we choose to label the alphabet

Desired properties

- It is a function of $p_1, \cdots, p_M$
- Let this uncertainty measure be

$$H(p_1, \cdots, p_M)$$

- Monotonicity. Let $f(M) = H(1/M, \cdots, 1/M)$. If $M < M'$, then

$$f(M) < f(M').$$

  Picking one person randomly from the classroom should result less possibility than picking a person randomly from the US.

- Additivity. Two independent RV X and Y, each uniformly distributed, alphabet size M and L. The uncertainty for the pair (X, Y), is ML. However, due to independence, when X is revealed, the uncertainty in Y should not be affected. This means

$$f(ML) - f(M) = f(L)$$

- Grouping rule (Problem 2.27 in Text). Dividing the outcomes into two, randomly choose one group, and then randomly pick an element from one group, does not change the number of possible outcomes.

Entropy

- The only function that satisfies the requirements is the entropy function

$$H(p_1, \cdots, p_M) = -\sum_{i=1}^{M} p_i \log_2 p_i$$

- General definition of entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \text{ bits}$$

- $0 \log 0 = 0$

- Uncertainty in a single random variable
- Can also be written as:

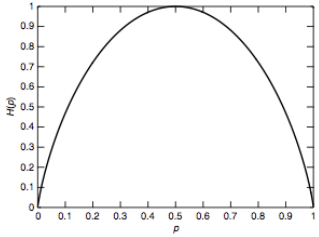
$$H(X) = \mathbb{E}\left\{\log \frac{1}{p(X)}\right\}$$

- Intuition: $H = \log(\#\text{of outcomes/states})$
- Entropy is a functional of $p(x)$
- Entropy is a lower bound on the number of bits need to represent a RV. E.g.: a RV that that has uniform distribution over 32 outcomes

Properties of entropy

- $H(X) \geq 0$
- Definition, for Bernoulli random variable, $X = 1$ w.p. p, $X = 0$ w.p. $1 - p$

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



  - Concave
  - Maximizes at $p = 1/2$

  Example: how to ask questions?

Joint entropy

- Extend the notion to a pair of discrete RVs (X, Y)
- Nothing new: can be considered as a single vector-valued RV
- Useful to measure dependence of two random variables

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

$$H(X,Y) = -\mathbb{E}\log p(X,Y)$$

Conditional Entropy

- Conditional entropy: entropy of a RV given another RV. If $(X, Y) \sim p(x, y)$

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

- Various ways of writing this

Chain rule for entropy

Entropy of a pair of RVs = entropy of one + conditional entropy of the other:

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

Proof:

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

Relative entropy

- Measure of distance between two distributions

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

- Also known as Kullback-Leibler distance in statistics: expected log-likelihood ratio

- A measure of inefficiency of assuming that distribution is q when the true distribution is p

- If we use distribution is q to construct code, we need $H(p) + D(p||q)$ bits on average to describe the RV

Mutual information

- Measure of the amount of information that one RV contains about another RV

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y))$$

- Reduction in the uncertainty of one random variable due to the knowledge of the other

- Relationship between entropy and mutual information

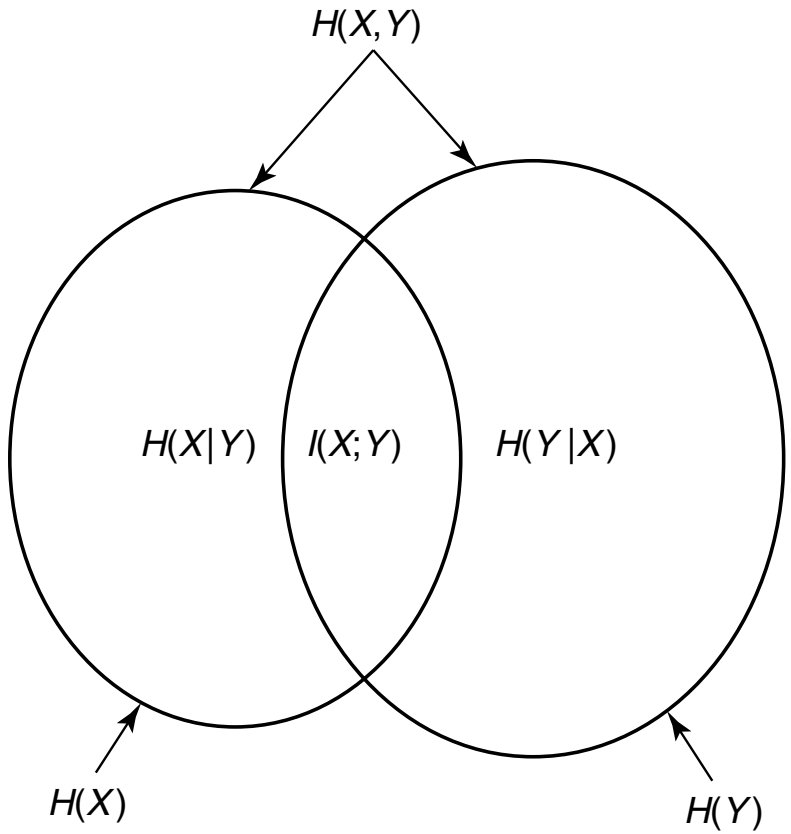
$$I(X;Y) = H(Y) - H(Y|X)$$

Proof:

- $I(X;Y) = H(Y) - H(Y|X)$
- $H(X,Y) = H(X) + H(Y|X) \rightarrow I(X;Y) = H(X) + H(Y) - H(X,Y)$
- $I(X;X) = H(X) - H(X|X) = H(X)$
  Entropy is "self-information"

Example: calculating mutual information

Vien diagram



$I(X;Y)$ is the intersection of information in X with information in Y

X: blood type

Y: chance for skin cancer

	A	**B**	**AB**	**O**
Very Low	1/8	1/16	1/32	1/32
Low	1/16	1/8	1/32	1/32
Medium	1/16	1/16	1/16	1/16
High	1/4	0	0	0

X: marginal (1/2, 1/4, 1/8, 1/8)

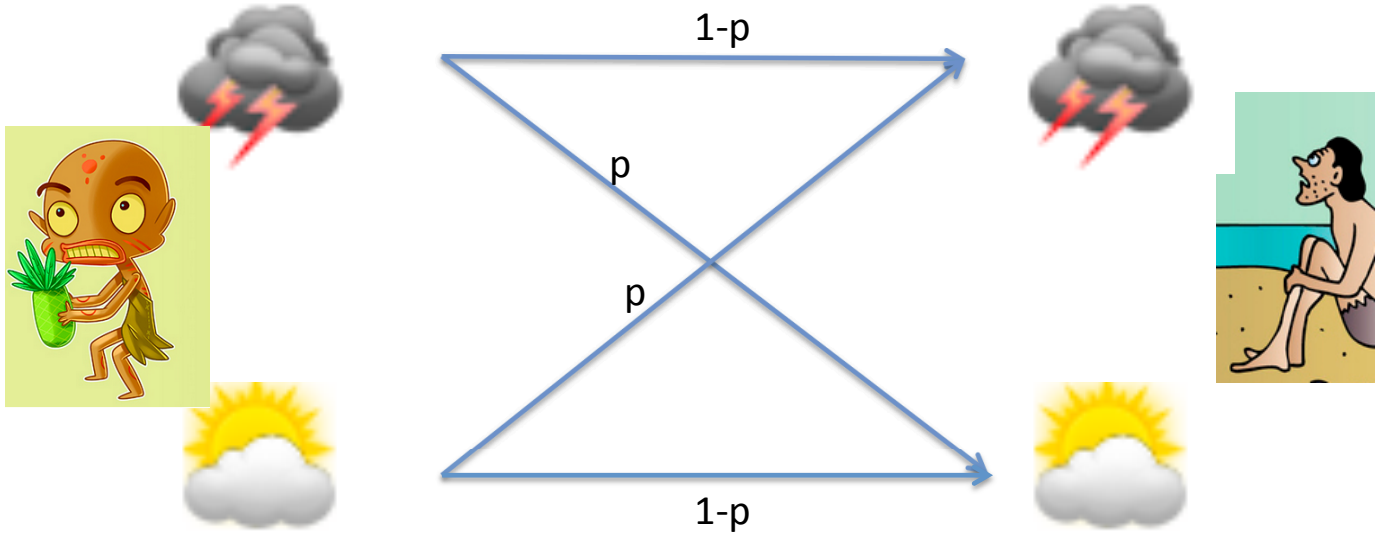
Y: marginal (1/4, 1/4, 1/4, 1/4)

$H(X) = 7/4$ bits $H(Y) = 2$ bits

Conditional entropy: $H(X|Y) = 11/8$ bits, $H(Y|X) = 13/8$ bits

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

Mutual information: $I(X; Y) = H(X) - H(X|Y) = 0.375$ bit

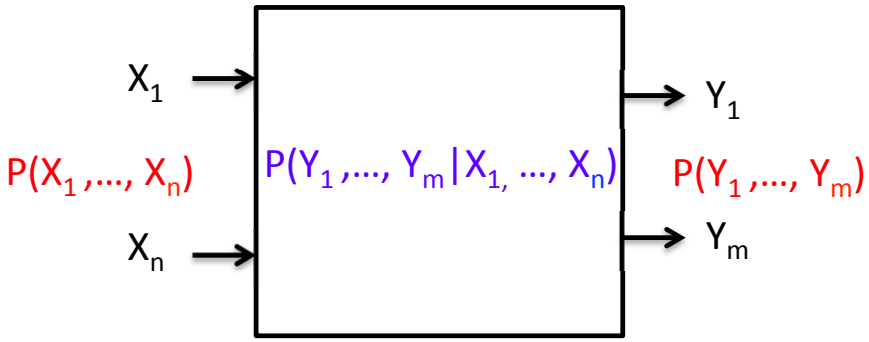

1-p
p
p
1-p

Summary

Entropy

$H(X)$

Mutual Information



$I(X_1, \ldots, X_n; Y_1, \ldots, Y_m)$