Information Theory

CS 4641

Charles Isbell
Chris Simpkins

Introduction

- Claude Shannon, A Mathematical Theory of Communication, 1948
- Questions:
  - What is the ultimate data compression?
  - What is the ultimate transmission rate of dat over a noisy channel?
- Answers:
  - Entropy
  - Relative Entropy
  - Mutual Information

Entropy

- A measure of the *uncertainty* in a random variable
- Tells you how much *information* you get from a trial
- Answers the question “what is the average length of the shortest description of X?”
- Discrete or continuous

Entropy Defined

$$H(X) = -\sum_{x \in \Omega} p(x) \log p(x)$$

or

$$H(X) = -E[\log p(x)]$$

- We also write *H(p)*
- Log is base 2, units are then bits
- 0 Log 0 = 0
- Functional of distribution of X. Depends only on probabilities, not values of X

Entropy of a Fair Coin

$$H(p) = -\sum_{x \in \Omega} p() \log p(x)$$

$$H(\frac{1}{2}) = -p(Head) \log p(Head) - p(Tail) \log p(Tail)$$

$$H(\frac{1}{2}) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2}$$

$$H(\frac{1}{2}) = 1$$

- We need a full bit to transmit information about one trial

Entropy of a Weighted Coin

$$H(p) = -\sum_{x \in \Omega} p() \log p(x)$$

$$H(\frac{3}{4}) = -p(Head) \log p(Head) - p(Tail) \log p(Tail)$$

$$H(\frac{3}{4}) = -\frac{3}{4} \log \frac{3}{4} - \frac{1}{4} \log \frac{1}{4}$$

$$H(\frac{3}{4}) = .811$$

- We need .811 bits to transmit information about one trial
- 82 bits suffices to represent 100 trials

Entropy of a 2-Headed Coin

$$H(p) = -\sum_{x \in \Omega} p() \log p(x)$$

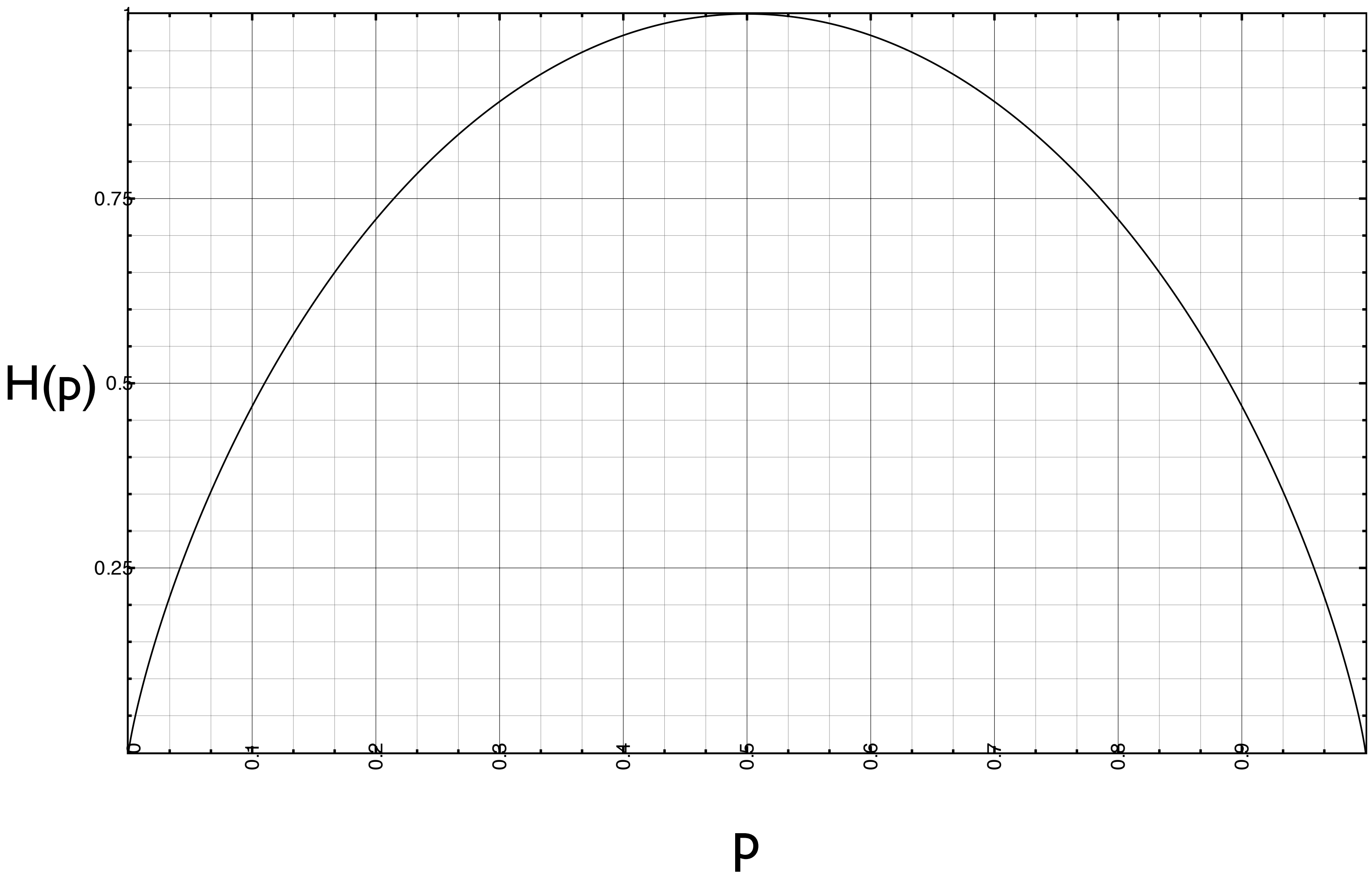
$$H(1) = -p(Head) \log p(Head) - p(Tail) \log p(Tail)$$

$$H(1) = -1 \log 1 - 0 \log 0$$

$$H(1) = 0$$

- 0 bits required to transmit one trial
- No information in a trial

Graph of H(p)



- H(p) is concave function of p
- 0 when p=0 or 1
- Max uncertainty when p = 1/2

Joint and Conditional Entropy

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

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

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

- Mirrors similar definitions for probabilities

Relative Entropy

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

$$D(p||q) = E_p[\log \frac{p(X)}{q(X)}]$$

$$D(p||q) = E_p[\log p(X) - \log q(X)]$$

- Also called *Kullback-Leibler* divergence
- Answers "given two variables, how similar are their distributions?"
- *D(p||q)* always non-negative, 0 only when *p=q*
- Non-symmetric, so not a true distance metric

Mutual Information

$$I(X;Y) = D(p(x,y)||p(x)p(y))$$

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

$$= E_p(x,y)[log \frac{p(X,Y)}{p(X)p(Y)}]$$

- Reduction in uncertainly of one variable given knowledge of the other
- If *I(X;Y) = 0* then *X* and *Y* are *independent*

Mutual Information and Entropy

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

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

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

$$I(X;Y) = I(Y;X)$$

$$I(X;X) = H(X)$$

Information Theory in Machine Learning

- Kolmogorov complexity K is shortest binary program that computes a string
  - If string drawn from distribution with entropy H, then $K \approx H$
- Decision trees
  - Information gain used to build short trees
- MIMIC
  - Mutual information used to build dependency trees