Basic Concepts in Information Theory

ChengXiang Zhai

Department of Computer Science

University of Illinois, Urbana-Champaign

Background on Information Theory

- Developed by Claude Shannon in the 1940s
- Maximizing the amount of information that can be transmitted over an imperfect communication channel
- Data compression (entropy)
- Transmission rate (channel capacity)

Claude E. Shannon: A Mathematical Theory of Communication, Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, 1948

Basic Concepts in Information Theory

- Entropy: Measuring uncertainty of a random variable
- Kullback-Leibler divergence: comparing two distributions
- Mutual Information: measuring the correlation of two random variables

Entropy: Motivation

- **Feature selection:**
  - If we use only a few words to classify docs, what kind of words should we use?
  - P(Topic| "computer"=1) vs p(Topic | "the"=1): which is more random?
- **Text compression:**
  - Some documents (less random) can be compressed more than others (more random)
  - Can we quantify the "compressibility"?
- **In general, given a random variable X following distribution p(X),**
  - How do we measure the "randomness" of X?
  - How do we design optimal coding for X?

Entropy: Definition

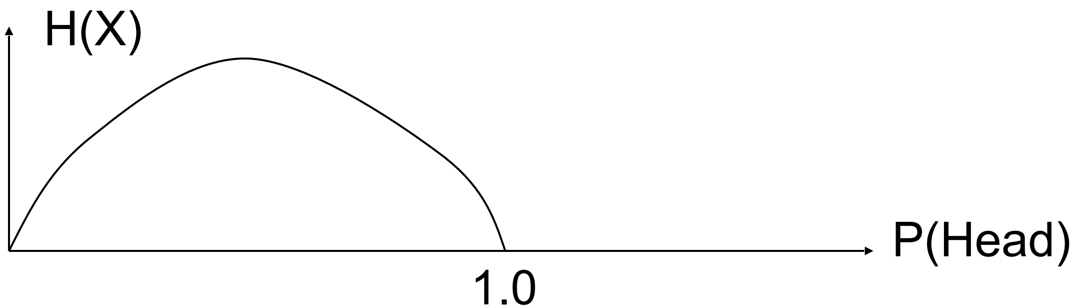
Entropy H(X) measures the uncertainty/randomness of random variable X

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

$$Define\ \ 0\log 0 = 0, \quad \log = \log_2$$

Example:

$$H(X) = \begin{cases} 1 & fair\ coin\ \ p(Head) = 0.5 \\ between\ 0\ and\ 1 & biased\ coin\ \ p(Head) = 0.8 \\ 0 & completely\ biased\ \ p(Head) = 1 \end{cases}$$



Entropy: Properties

- **Minimum value of H(X): 0**
  - **What kind of X has the minimum entropy?**
- **Maximum value of H(X): log M, where M is the number of possible values for X**
  - **What kind of X has the maximum entropy?**
- **Related to coding**

$$H(X) = -\sum_{x \in \Omega} p(x)\log_2 p(x)$$

$$= \sum_{x \in \Omega} p(x)\log_2 \frac{1}{p(x)}$$

$$= E\left(\log_2 \frac{1}{p(x)}\right)$$

$$\text{"}\mathit{Information\ of\ x}\text{"} = \text{"}\mathit{\#bits\ to\ code\ x}\text{"} = -\log p(x) \qquad H(X) = E_p[-\log p(x)]$$

Interpretations of H(X)

- Measures the “amount of information” in X
  - Think of each value of X as a “message”
  - Think of X as a random experiment (20 questions)
- Minimum average number of bits to compress values of X
  - The more random X is, the harder to compress

A fair coin has the maximum information, and is hardest to compress
A biased coin has some information, and can be compressed to <1 bit <u>on aver</u>
A completely biased coin has no information, and needs only 0 bit

$$\text{"}Information\ of\ x\text{"} = \text{"}\#bits\ to\ code\ x\text{"} = -\log p(x) \quad H(X) = E_p[-\log p(x)]$$

Conditional Entropy

- **The conditional entropy of a random variable Y given another X, expresses how much extra information one still needs to supply on average to communicate Y given that the other party knows X**

$$H(Y \mid X) = \sum_{x \in \Omega_X} p(x) H(Y \mid X = x)$$

$$= -\sum_{x \in \Omega_X} p(x) \sum_{y \in \Omega_Y} p(y \mid x) \log p(y \mid x)$$

$$= -\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \log p(y \mid x) \; = -E\left(\log p(Y \mid X)\right)$$

- **H(Topic| "computer") vs. H(Topic | "the")?**

Cross Entropy H(p,q)

What if we encode X with a code optimized for a wrong distribution q?

Expected # of bits=? $\boxed{H(p,q) = E_p[-\log q(x)] = -\sum_{x \in \Omega} p(x) \log q(x)}$

Intuitively, H(p,q) ≥ H(p), and mathematically,

$$H(p,q) - H(p) = \sum_{x \in \Omega} p(x)[-\log \frac{q(x)}{p(x)}]$$

$$\geq -\log \sum_{x \in \Omega} [p(x) \frac{q(x)}{p(x)}] = 0$$

By Jensen's inequality: $\sum_i p_i f(x_i) \geq f(\sum_i p_i x_i)$

where, f is a convex function, and $\sum_i p_i = 1$

Kullback-Leibler Divergence D(p||q)

What if we encode X with a code optimized for a wrong distribution q?

How many bits would we waste?

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

Relative entropy

Properties:

- $D(p\|q) \geq 0$
- $D(p\|q) \neq D(q\|p)$
- $D(p\|q) = 0$ iff $p=q$

→ **KL-divergence is often used to measure the distance between two distributions**

Interpretation:

-Fix p, D(p||q) and H(p,q) vary in the same way

-If p is an empirical distribution, minimize D(p||q) or H(p,q) is equivalent to maximizing likelihood

Cross Entropy, KL-Div, and Likelihood

$$\text{Random Var}: X \in \{x_1, \ldots, x_n\} \quad \text{prob. given by a model}: \{p(X = x_i)\}$$

$$\text{Data Sample (i.i.d)}: \quad Y = (y_1 y_2 \ldots y_N),\ y_i \in \{x_1, \ldots, x_n\}$$

Example: X ∈{"H","T"}
Y=(HHTTH)

$$\tilde{p}(X = "H") = \frac{c("H", Y)}{5} = 3/5$$

$$\text{Empirical distribution}: \tilde{p}(X = x_i) = \frac{count(x_i, Y)}{N} = \frac{\sum_{j=1}^{N} \delta(y_j, x_i)}{N}$$

$$\delta(y, x) = \begin{cases} 1 & if\ x = y \\ 0 & otherwise \end{cases}$$

$$\text{loglikelihood}: \log L(Y) = \sum_{j=1}^{N} \log p(X = y_j) = \sum_{i=1}^{n} count(x_i, Y) \log p(X = x_i) = N \sum_{i=1}^{n} \tilde{p}(x_i) \log p(x_i)$$

$$\frac{1}{N} \log L(Y) = -H(\tilde{p}, p) = -D(\tilde{p} \| p) - H(\tilde{p})$$

$$\textit{Fix the data} \Rightarrow \textit{fix}\ Y, \tilde{p}$$

$$p^* = \arg\max_p \frac{1}{N} \log L(Y) = \arg\min_p H(\tilde{p}, p) = \arg\min_p D(\tilde{p} \| p) = \arg\min_p 2^{-\frac{1}{N}\log L(Y)}$$

Equivalent criteria for selecting/evaluating a model
Perplexity(p)

Mutual Information I(X;Y)

Comparing two distributions: p(x,y) vs p(x)p(y)

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

Properties: I(X;Y)≥0; I(X;Y)=I(Y;X); I(X;Y)=0 iff X & Y are independent

Interpretations:

- Measures how much reduction in uncertainty of X given info. about Y
- Measures correlation between X and Y
- Related to the “channel capacity” in information theory

Examples:

I(Topic; “computer”) vs. I(Topic; “the”)?

I(“computer”, “program”) vs I(“computer”, “baseball”)?

What You Should Know

- Information theory concepts: entropy, cross entropy, relative entropy, conditional entropy, KL-div., mutual information
  - Know their definitions, how to compute them
  - Know how to interpret them
  - Know their relationships