Information Theory

A Tutorial for Machine Learning

Shenghao Yang

03 April 2017

Shenzhen Research Institute of Big Data
The Chinese University of Hong Kong, Shenzhen

What is information

- Information is about uncertainty.
- Entropy is a measure of the uncertainty of a random variable, and arises naturally as the fundamental limits of source coding.
- Mutual information measures certain dependence of two random variables, and arises naturally as the fundamental limits of channel coding.

Entropy

Entropy

- Let X be a discrete random variable with a finite alphabet $\mathcal{X}$.
- Let distribution $p(x) \triangleq \Pr\{X = x\}$, $x \in \mathcal{X}$.

Definition

The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_x p(x) \log p(x).$$

Remark

1. The summation is over the support of X.
2. The log is to the base 2 and the unit of entropy is *bit*.
3. $H(X)$ depends only on $p(x)$, not on the actual values of x—entropy is independent of the alphabet $\mathcal{X}$.

Example

Consider a random variable that has a uniform distribution over 32 outcomes. The entropy of this random variable is 5 bits.

Example

Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right).$$

The entropy of the horse race is 2 bits.

Other Forms

- $-\log p(x)$ is called self-information.
- Expectation form $H(X) = \mathbb{E}[-\log(p(X)]$.
- Binary entropy function: $H(p) = -p\log p - (1-p)\log(1-p)$

Properties

- $H(X) \geq 0$ where equality holds iff X is a constant.
- $H(X) \leq \log |\mathcal{X}|$ where $\mathcal{X}$ is the alphabet of X. The equality holds iff X is uniformly distributed on $\mathcal{X}$.

Conditional Entropy

- For random variables X and Y, the *conditional entropy* $H(Y|X)$ is defined as

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

- Denote

$$H(Y|X=x) = H(p_{Y|X}(\cdot|x)) = -\sum_{y} p(y|x) \log p(y|x).$$

- We can write

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

- In other words, the conditional entropy is the expectation of the entropy of the conditional distribution of Y given $X = x$.

Basic Properties

- $H(Y|X) \geq 0$ with equality iff Y is a function of X (over the support of X).
- (Chain rule) $H(X, Y) = H(X) + H(Y|X)$.
- $H(Y|X) \leq H(Y)$ with equality iff X and Y are independent. In other words, conditioning reduces entropy.

Mutual Information

Mutual Information

Definition

The *mutual information* between random variables X and Y is defined as

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \mathbb{E} \log \frac{p(X,Y)}{p(X)p(Y)}.$$

Remark

1. $I(X;Y)$ is symmetrical in X and Y.
2. $I(X;X) = H(X)$: observing X can get all the information of X.
3. $I(X;Y)$ only depends on the joint distribution $p_{X,Y}$, so we also write $I(X;Y) = I(p_{X,Y})$.

Relations

- We have the following equalities:

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

- If the alphabets are not finite, the above equalities hold provided that all the entropies and conditional entropies are finite.

- $I(X,Y) \geq 0$, with equality if and only if X and Y are independent.
- $I(X,Y|Z) \geq 0$, with equality if and only if X and Y are independent given Z.

Example

Noiseless binary channel.

Example

Noisy four-symbol channel.

Example

Binary symmetric channel.

Relative Entropy

Relative Entropy

Definition

The *relative entropy* (*information divergence* or *Kullback-Leibler distance*) between two probability mass function $p(x)$ and $q(x)$ is defined as

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

Remark

- $I(X;Y) = D(p(x,y)\|p(x)p(y))$.
- $D(p\|q) \geq 0$ with equality iff $p = q$.

Convexity

- $D(p\|q)$ is convex in the pair (p, q), which implies
- $H(p)$ is a concave function of p, and
- $I(X;Y)$ is 1) a concave function of $p(x)$ for fixed $p(y|x)$ and is 2) a convex function of $p(y|x)$ for fixed $p(x)$.

Reading

Further Reading

- T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 2006.