CS258: Information Theory

Fan Cheng
Shanghai Jiao Tong University

http://www.cs.sjtu.edu.cn/~chengfan/
chengfan@sjtu.edu.cn
Spring, 2020

Outline

- Entropy
- Relative entropy
- Mutual information
- Information inequality

Relative Entropy

Relative entropy: a measure of the distance between two distributions

- Probability is not linear, but log function can alleviate it

The relative entropy or Kullback–Leibler (KL) distance between two probability mass functions $p(x)$ and $q(x)$ over the alphabet $\mathcal{X}$ is defined as

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

$$= E_p \log \frac{p(X)}{q(X)}$$

- $0 \log \frac{0}{0} = 0, 0 \log \frac{0}{q} = 0,\ p \log \frac{p}{0} = \infty$
- If there exists $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p||q) = \infty$
- $D(p||q) \geq 0$ (Show it later)
- $D(p||q) = E_p(-\log q(x)) - E_p(-\log p(x)) = E_p(-\log q(x)) - H(p)$

Relative Entropy Not Metric

A metric (测度) $d: X, Y \to R^+$ between two distributions should satisfy

- $d(X,Y) \geq 0$
- $d(X,Y) = d(Y,X)$
- $d(X,Y) = 0$ if and only if $X = Y$
- $d(X,Y) + d(Y,Z) \geq d(X,Z)$

- The Euclidean distance is a metric
- **KL distance is not a metric**
  - $\boldsymbol{D(p||p) = 0}$
  - $D(p||q) \neq D(q||p)$

    $p = (1/2, 1/2), q = (1/4,3/4), D(p||q) = 0.2, D(q||p) = 0.18$
- The variational distance between p and q is denoted as

$$V(p,q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|$$

- Pinsker's inequality

$$D(p||q) \geq \frac{1}{2\ln 2} V^2(p,q)$$

Mutual Information

Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X;\ Y)$ is the relative entropy between the joint distribution $p(x, y)$ and the product distribution $p(x)p(y)$:

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$
$$= D(p(x,y)||p(x)p(y))$$
$$= E_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}$$

- $I(X;Y) = I(Y;X)$
- $I(X;X) = H(X)$
- X and Y are independent
  $I(X;Y) = 0$
- Common mistakes:
  - $I(X,Y)$ $H(X;Y)$

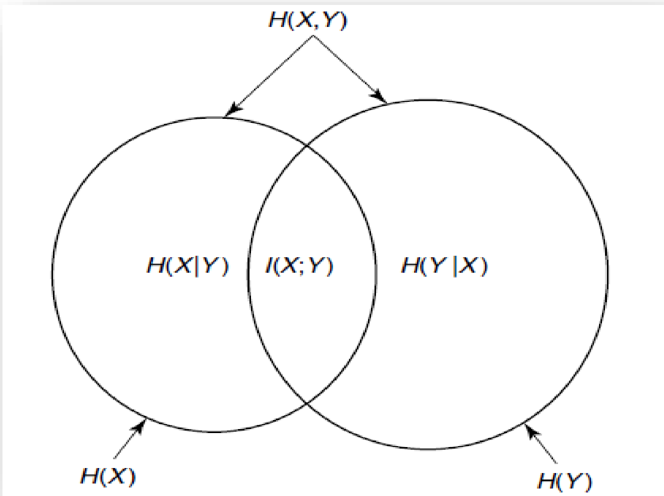
$Y \backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

$I(X;Y)$?

What is the relationship of $H(X), H(Y), I(X;Y), H(X|Y), H(Y|X)$?

Mutual Information and Entropy

- Venn diagram for $H(X,Y), H(X), H(Y), I(X;Y)$



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

Proof sketch

- Fact $p(X,Y) = p(X)p(Y|X) = p(Y)p(X|Y)$
- Take $\log()$ at each side
- Take expectation E at both sides: $E(X_1 + X_2) = E(X_1) + E(X_2)$
- To prove $I(X;Y) = H(X) + H(Y) - H(X,Y)$

$$\log \frac{p(X,Y)}{p(X)p(Y)} = -\log p(X) - \log p(Y) + \log p(X,Y)$$

The point of view to look at $p(X,Y)$ determine all the equalities.

Chain Rule for Entropy

For a collection of random variables $X_1, X_2, \ldots, X_n$

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2, \ldots, x_n|x_1) = \cdots$$
$$= p(x_1)p(x_2|x_1) \ldots p(x_n|x_1, \ldots, x_{n-1})$$

Take expectations E

> Chain rule for entropy: Let $X_1, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then
> $$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, \ldots, X_1)$$

- If $X_1, X_2, \ldots, X_n$ are independent,

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i)$$

- For two random variables X, Y, if X and Y are independent, then

$$H(X, Y) = H(X) + H(Y)$$
$$I(X; Y) = H(X) + H(Y) - H(X, Y) = 0$$

and vice versa (prove later).

Chain Rule for Information

The conditional mutual information of random variables X and Y given Z is defined by

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

Chain rule for information

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1)$$

Proof sketch

- $I(X_1, X_2, \dots, X_n; Y) = H(X_1, \dots, X_n) - H(X_1, \dots, X_n | Y)$
- Chain rule for $H(X_1, \dots, X_n), H(X_1, \dots, X_n | Y)$, respectively
  - $H(X_1, \dots, X_n)$
  - $H(X_1, \dots, X_n | Y)$

Conditional Relative Entropy

For joint probability mass functions $p(x, y)$ and $q(x, y)$, the conditional relative entropy $D(p(y|x)||q(y|x))$ is the average of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$. More precisely,

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

$$= \sum_x \sum_y p(x)p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

$$= E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}$$

Chain rule for relative entropy

$$D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

By definition

$$D(p(x, y)||q(x, y)) = \sum_x \sum_y p(x, y) \log \frac{p(x,y)}{q(x,y)} = \sum_x \sum_y p(x, y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)}$$

$$= \sum_x \sum_y p(x, y) \left(\log \frac{p(x)}{q(x)} + \log \frac{p(y|x)}{q(y|x)}\right)$$

Probability distribution
$p(x_1, x_2, \dots, x_n)$

Information quantities
$H(X_1, X_2, \dots, X_n), H(\ \ |\ \ ), I(;)$

Law on probability
$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2|x_1)\dots p(x_n|\dots)$

Expectation: E

Linear relations on
$H(\ \ ), I(;)$; e.g. Chain rule

$D(p||q) \geq 0$

Information inequality: Let $p(x), q(x), x \in X$, be two probability mass functions. Then

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

$$D(p||q) = \sum p \log \frac{p}{q}$$

Two proofs with hints:

- By convexity/concavity

$$-D(p||q) = \sum p \log \frac{q}{p} \leq \log \sum p \frac{q}{p} = \log \sum q \leq \log 1 = 0$$

- Using $\log x \leq x - 1$ when $x > 0$

$$-D(p||q) = \sum p \log \frac{q}{p} \leq \sum p \left(\frac{q}{p} - 1\right) = \sum q - \sum p \leq 0$$

$D(p||q) \geq 0$

Corollary: **(Homework 2)**

- $D(p||q) = 0$, if and only if $p(x) = q(x)$
- $I(X;Y) \geq 0$, with equality if and only if X and Y are independent
- $D(p(y|x)||q(y|x)) \geq 0$ with equality if and only if $p(y|x) = q(y|x)$ for all y and x such that $p(x) > 0$
- $I(X;Y|Z) \geq 0$ with equality if and only if X and Y are conditionally independent given Z
- Let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform probability mass function over X, and let $p(x)$ be the probability mass function for X. Then

$$0 \leq D(p||u) = \log|\mathcal{X}| - H(X)$$

- (Conditioning reduces entropy) (Information can't hurt)

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

with equality if and only if X and Y are independent

Summary

The materials of this lecture are related to

- The textbook of T. Cover: 2.3, 2.4., 2.5, 2.6