CS258: Information Theory

Fan Cheng

Spring, 2018. chengfan@sjtu.edu.cn

Recap: Error Correcting

- Channel model: BSC and its channel capacity
- Repetition code: encoder, decoder, information rate
- (7, 4) Hamming code: encoder, decoder, information rate

- Information content
- Shannon source coding theorem
- Typical set
- Fundamental tools

A mathematical brain-teaser

A mathematical brain-teaser

Oddball Problem

A mathematical brain-teaser

Oddball Problem

- Given 12 balls, all equal in weight except for one that is either heavier or lighter.

A mathematical brain-teaser

Oddball Problem

- Given 12 balls, all equal in weight except for one that is either heavier or lighter.
- You are also given a two-pan balance to use. In each use of the balance, there are three possible outcomes: equal, heavier, or lighter.

A mathematical brain-teaser

Oddball Problem

- Given 12 balls, all equal in weight except for one that is either heavier or lighter.
- You are also given a two-pan balance to use. In each use of the balance, there are three possible outcomes: equal, heavier, or lighter.
- Design a strategy to determine which is the odd ball and whether it is heavier or lighter in as few uses of the balance as possible.

(a) How can one measure information?

(a) How can one measure information?

(b) When you have identified the odd ball and whether it is heavy or light, how much information have you gained?

Questions

(a) How can one measure information?

(b) When you have identified the odd ball and whether it is heavy or light, how much information have you gained?

(c) Once you have designed a strategy, draw a tree showing, for each of the possible outcomes of a weighing, what weighing you perform next. At each node in the tree, how much information have the outcomes so far given you, and how much information remains to be gained?

Questions

(a) How can one measure information?
(b) When you have identified the odd ball and whether it is heavy or light, how much information have you gained?
(c) Once you have designed a strategy, draw a tree showing, for each of the possible outcomes of a weighing, what weighing you perform next. At each node in the tree, how much information have the outcomes so far given you, and how much information remains to be gained?
(d) How much information is gained on the first step of the weighing problem if 6 balls are weighed against the other 6? How much is gained if 4 are weighed against 4 on the first step, leaving out 4 balls?

Ensemble and Information content

An ensemble X is a triple $(x; \mathcal{A}_X; \mathcal{P}_X)$, where the outcome x is the value of a random variable, which takes on one of a set of possible values, $\mathcal{A}_X = \{a_1, a_2, ..., a_i, ..., a_I\}$, having probabilities $\mathcal{P}_X = \{p_1, p_2, ..., p_I\}$, with $P(x = a_i) = p_i, p_i \geq 0$ and $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$.

An ensemble X is a triple $(x; \mathcal{A}_X; \mathcal{P}_X)$, where the outcome x is the value of a random variable, which takes on one of a set of possible values, $\mathcal{A}_X = \{a_1, a_2, ..., a_i, ..., a_I\}$, having probabilities $\mathcal{P}_X = \{p_1, p_2, ..., p_I\}$, with $P(x = a_i) = p_i, p_i \geq 0$ and $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$.

Shannon Information Content

The Shannon information content of the outcome $x = a_i$ is

$$h(x = a_i) = \log_2 \frac{1}{p_i}$$

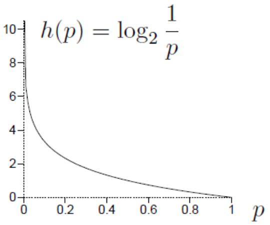
Ensemble and Information content

An ensemble X is a triple $(x; \mathcal{A}_X; \mathcal{P}_X)$, where the outcome x is the value of a random variable, which takes on one of a set of possible values, $\mathcal{A}_X = \{a_1, a_2, ..., a_i, ..., a_I\}$, having probabilities $\mathcal{P}_X = \{p_1, p_2, ..., p_I\}$, with $P(x = a_i) = p_i, p_i \geq 0$ and $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$.

Shannon Information Content

The Shannon information content of the outcome $x = a_i$ is

$$h(x = a_i) = \log_2 \frac{1}{p_i}$$



Guessing games

The game ‘sixty-three’. What’s the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

The game 'sixty-three'. What's the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

1: is $x \geq 32$?

2: is $x \bmod 32 \geq 16$?

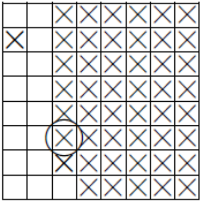
3: is $x \bmod 16 \geq 8$?

4: is $x \bmod 8 \geq 4$?

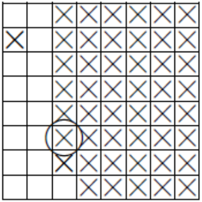
5: is $x \bmod 4 \geq 2$?

6: is $x \bmod 2 = 1$?

The game of submarine

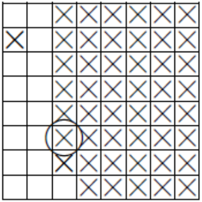


The game of submarine



The entropy is: $H(X) = \log_2 64 = 6$

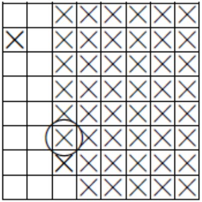
The game of submarine



The entropy is: $H(X) = \log_2 64 = 6$
If we hit the submarine when there are n squares left to choose from, then the total information gained is:

$$\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \ldots + \log_2 \frac{n+1}{n} + \log_2 \frac{n}{1} = log_2 64 = 6$$

The game of submarine



The entropy is: $H(X) = \log_2 64 = 6$

If we hit the submarine when there are n squares left to choose from, then the total information gained is:

$$\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \ldots + \log_2 \frac{n+1}{n} + \log_2 \frac{n}{1} = log_2 64 = 6$$

Meaning of Shannon information content

Shannon information content measures the length of a binary file that encodes x.

Lossy Compression

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

Lossy Compression

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = log_2|\mathcal{A}_X|$$

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = log_2|\mathcal{A}_X|$$

Idea: Discard some outcomes to reduce the raw bit content

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = log_2|\mathcal{A}_X|$$

Idea: Discard some outcomes to reduce the raw bit content

The smallest δ-sufficient subset S_δ is the smallest subset of $\mathcal{A}_X$ satisfying

$$P(x \in S_\delta) \geq 1 - \delta$$

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = log_2|\mathcal{A}_X|$$

Idea: Discard some outcomes to reduce the raw bit content

The smallest δ-sufficient subset S_δ is the smallest subset of $\mathcal{A}_X$ satisfying

$$P(x \in S_\delta) \geq 1 - \delta$$

The essential bit content of X is

$$H_\delta(X) = \log_2 |S_\delta|$$

Lossy Compression

One way of measuring the information content of a random variable is simply to count the number of possible outcomes, $|\mathcal{A}_X|$

The raw bit content of X is

$$H_0(X) = log_2|\mathcal{A}_X|$$

Idea: Discard some outcomes to reduce the raw bit content

The smallest δ-sufficient subset S_δ is the smallest subset of $\mathcal{A}_X$ satisfying

$$P(x \in S_\delta) \geq 1 - \delta$$

The essential bit content of X is

$$H_\delta(X) = \log_2 |S_\delta|$$

Denote by X^N the ensemble $(X_1, X_2, ..., X_N)$, where X_i's are independent identically distributed random variables. What is $H_\delta(X^N)$?

Shannon's source coding theorem

Theorem

Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$|\frac{1}{N}H_\delta(X^N) - H| < \epsilon$$

Theorem

Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$|\frac{1}{N}H_\delta(X^N) - H| < \epsilon$$

- Typicality: Law of large numbers

Theorem

Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,

$$|\frac{1}{N} H_\delta(X^N) - H| < \epsilon$$

- Typicality: Law of large numbers
- Some useful fundamental inequalities

Typicality

By law of large numbers, the probability of a typical string $x \in \mathcal{A}_X^N$ is

$$P(x) = P(x_1)P(x_2)...P(x_N) \simeq p_1^{p_1 N} p_2^{p_2 N} ... p_I^{p_I N}$$

Typicality

By law of large numbers, the probability of a typical string $x \in \mathcal{A}_X^N$ is

$$P(x) = P(x_1)P(x_2)...P(x_N) \simeq p_1^{p_1 N} p_2^{p_2 N} ... p_I^{p_I N}$$

The information content is

$$\log_2 \frac{1}{P(x)} \simeq N \sum_i p_i \log_2 \frac{1}{p_i} = NH$$

Typicality

By law of large numbers, the probability of a typical string $x \in \mathcal{A}_X^N$ is

$$P(x) = P(x_1)P(x_2)...P(x_N) \simeq p_1^{p_1 N} p_2^{p_2 N} ... p_I^{p_I N}$$

The information content is

$$\log_2 \frac{1}{P(x)} \simeq N \sum_i p_i \log_2 \frac{1}{p_i} = NH$$

Typical set

$$T_{N\beta} := \{x \in \mathcal{A}_X^N : |\frac{1}{N} \log_2 \frac{1}{P(x)} - H| < \beta\}$$

'Asymptotic equipartition' principle (AEP). With N sufficiently large, the outcome $x = (x_1, x_2, ..., x_N)$ is almost certain to belong to a subset of $\mathcal{A}_X^N$ having only $2^{NH(X)}$ members, each having probability 'close to' $2^{-NH(X)}$.

Chebyshev's inequality 1

Let t be a non-negative real random variable, and let α be a positive real number. Then

$$P(t \geq \alpha) \leq \frac{\bar{t}}{\alpha},$$

where $\bar{t}$ is the mean of t.

$P(t \geq \alpha) = \sum_{t \geq \alpha} P(t)$. We multiply each term by $t/\alpha \geq 1$ and obtain: $P(t \geq \alpha) \leq \sum_{t \geq \alpha} P(t) t/\alpha$. We add the (non-negative) missing terms and obtain: $P(t \geq \alpha) \leq \sum_{t} P(t) t/\alpha = \bar{t}/\alpha$

Chebyshev's inequality 1

Let t be a non-negative real random variable, and let α be a positive real number. Then

$$P(t \geq \alpha) \leq \frac{\bar{t}}{\alpha},$$

where $\bar{t}$ is the mean of t.

$P(t \geq \alpha) = \sum_{t \geq \alpha} P(t)$. We multiply each term by $t/\alpha \geq 1$ and obtain: $P(t \geq \alpha) \leq \sum_{t \geq \alpha} P(t)t/\alpha$. We add the (non-negative) missing terms and obtain: $P(t \geq \alpha) \leq \sum_t P(t)t/\alpha = \bar{t}/\alpha$

Chebyshev's inequality 2

Let x be a random variable, and let α be a positive real number. Then

$$P((x - \bar{x})^2 \geq \alpha) \leq \sigma_x^2/\alpha$$

Take $t = (x - \bar{x})^2$.

Take x to be the average of N independent random variables $h_1, h_2, \ldots, h_N$, having common mean $\bar{h}$ and common variance σ_h^2. Then

$$P((x - \bar{h})^2 \geq \alpha) \leq \sigma_h^2 / \alpha N$$

Take $\bar{x} = \bar{h}$ and $\sigma_x^2 = \sigma_h^2 / N$.

Proof of Source Coding Theorem

- $\frac{1}{N} H_\delta(X^N) < H + \epsilon$
- $\frac{1}{N} H_\delta(X^N) > H - \epsilon$

Reading and Exercise

Reading: Ch. 4 (MacKay)

Exercise

4.9, 4.10, 4.11, 4.12