COMP2610 – Information Theory

Lecture 11: Entropy and Coding

Australian National University

27 August 2018

Brief Recap of Course (Last 6 Weeks)

- How can we quantify information?
  - Basic Definitions and Key Concepts
  - Probability, Entropy & Information
- How can we make good guesses?
  - Probabilistic Inference
  -
- Ho
  -
  - Source Coding Theorem, Kraft Inequality
  - Block, Huffman, and Lempel-Ziv Coding
- How much noise can we correct and how?
  - Noisy-Channel Coding
  - Repetition Codes, Hamming Codes
- What is randomness?
  - Kolmogorov Complexity
  - Algorithmic Information Theory

Brief Overview of Course (Next 6 Weeks)

- How can we quantify information?
  - Basic Definitions and Key Concepts
  - Probability, Entropy & Information
- How can we make good guesses?
  - Probabilistic Inference
- Ho
  - Source Coding Theorem, Kraft Inequality
  - Block, Huffman, and Lempel-Ziv Coding
- How much noise can we correct and how?
  - Noisy-Channel Coding
  - Repetition Codes, Hamming Codes
- What is randomness?
  - Kolmogorov Complexity
  - Algorithmic Information Theory

Brief Overview of Course (Next 6 Weeks)

- How can we quantify information?
  - Basic Definitions and Key Concepts
  - Probability, Entropy & Information
- How can we make good guesses?
  - Probabilistic Inference
- Ho
  - Source Coding Theorem, Kraft Inequality
  - Block, Huffman, and Lempel-Ziv Coding
- How much noise can we correct and how?
  - Noisy-Channel Coding
  - Repetition Codes, Hamming Codes
- What is randomness?
  - Kolmogorov Complexity
  - Algorithmic Information Theory

This time

Basic goal of compression

Key concepts: codes and their types, raw bit content, essential bit content

Informal statement of source coding theorem

1. Introduction
   - Overview
   - What is Compression?
   - A Communication Game
   - What's the best we can do?

2. Form
   - En
   - Defining Codes

3. Formalising Compression
   - Reliability vs. Size
   - Key Result: The Source Coding Theorem

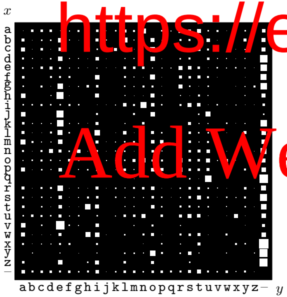
What is Compression?

Cn y rd ths mssg wtht ny vwls?

What is Compression?

Cn y rd ths mssg wtht ny vwls?

It is not too difficult to read as there is redundancy in English text.
(Estimates of 1-1.5 bits per character, compared to $\log$ 26 4.7)



(a) $P(y \mid x)$

- to be
- Th
  co
- Co
  differences in relative probability of symbols or blocks of symbols

Compression in a Nutshell

Compre

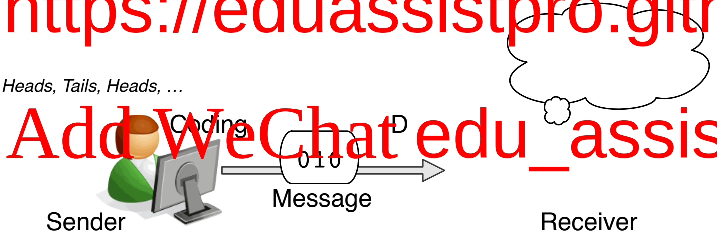
Data com

messag

A General Communication Game

Imagine the following game between Sender & Receiver:

- Sender & Receiver agree on code for each outcome ahead of time (e.g., 0 for *Heads*; 1 for *Tails*)
- Sender observes outcomes then codes and sends message
- Receiver decodes message and recovers outcome sequence



Goal: Want small messages on average when outcomes are from a fixed, known, but uncertain source (e.g., coin flips with known bias)

Sneak peek: source coding theorem

Consider a coin with $P(\mathit{Heads}) = 0.9$. If we want perfect transmission:

- Co
- Co

Not very interesting!

Sneak peek: source coding theorem

Consider a coin with $P(Heads) = 0.9$. If we want perfect transmission:

- Coding single outcomes requires 1 bit/outcome
- Co

Not very in

Things get interesting if we:

- accept errors in transmission
- allow variable length messages

Sneak peek: source coding theorem

Consider a coin with $P(Heads) = 0.9$. If we want perfect transmission:

- Coding single outcomes requires 1 bit/outcome
- Co

Not very in

Things get interesting if we:

- **accept errors in transmission** (this w
- allow variable length messages (next week)

Sneak peek: source coding theorem

If we are happy to fail on up to 2% of the sequences we can ignore any sequence of 10 outcomes with more than 3 tails

Why? T

Sneak peek: source coding theorem

If we are happy to fail on up to 2% of the sequences we can ignore any sequence of 10 outcomes with more than 3 tails

Why? T

There are

So, we can just code those, and **ignore** th

- Coding 10 outcomes with 2% failure doable wi bits/outcome
- *Smallest bits/outcome needed for 10,000 outcome sequences?*

Generalisation: Source Coding Theorem

What happens when we generalise to arbitrary error probability, and sequence size?

Generalisation: Source Coding Theorem

What happens when we generalise to arbitrary error probability, and sequence size?

Source

If: you wan
degree of r

Then: the average number of bits per outcome you w hly
equal to the entropy of that source.

Generalisation: Source Coding Theorem

What happens when we generalise to arbitrary error probability, and sequence size?

Source

If: you wan

degree of r

Then: the average number of bits per outcome you w hly equal to the entropy of that source.

To define: “Uniformly code”, “large sequences”, “degree of reliability”, “average number of bits per outcome”, “roughly equal”

1 Introduction
- Overview
- What is Compression?
- A Communication Game
- What's the best we can do?

2 Form
- En
- Defining Codes

3 Formalising Compression
- Reliability vs. Size
- Key Result: The Source Coding Theorem

Entropy and Information: Recap

Ensemble

An **ensemble** X is a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$; x is a **random variable** taking **values** in $\mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$ with **probabilities** $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$

Entropy and Information: Recap

Ensemble

An **ensemble** X is a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$; x is a **random variable** taking **values** in $\mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$ with **probabilities** $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$

Informa

The **info** X) is

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

Entropy and Information: Recap

Ensemble

An **ensemble** X is a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$; x is a **random variable** taking **values** in $\mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$ with **probabilities** $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$

Informa

The **info** X) is

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

Entropy

The **entropy** of an ensemble X is the average information

$$H(X) = \mathbb{E}[h(x)] = \sum_i p_i h(a_i) = \sum_i p_i \log_2 \frac{1}{p_i}$$

What is a Code?

A source code is a process for assigning names to outcomes. The names are typically expressed by strings of binary symbols.

We will denote the set of all finite binary strings by

$$+ \stackrel{\text{def}}{}$$

Source

Given an e : $\mathcal{A}_X$, **rce code** for X. The number of symbols in $c(x)$ is the **l** or x.

The **extension** of c is defined by $c(x_1 \ldots x$

What is a Code?

A source code is a process for assigning names to outcomes. The names are typically expressed by strings of binary symbols.

We will denote the set of all finite binary strings by

$$+ \quad \text{def}$$

Source

Given an e : $\mathcal{A}_X$, **rce code** for X. The number of symbols in $c(x)$ is the **l** or x.
The **extension** of c is defined by $c(x_1 \ldots x$

Example:

- The code c names outcomes from $\mathcal{A}_X = \{\mathtt{r}, \mathtt{g}, \mathtt{b}\}$ by $c(\mathtt{r}) = \mathtt{00}$, $c(\mathtt{g}) = \mathtt{10}$, $c(\mathtt{b}) = \mathtt{11}$
- The length of the codeword for each outcome is 2.
- The extension of c gives $c(\mathtt{rgrb}) = \mathtt{00100011}$

Types of Codes

Let X be an ensemble and $c : \mathcal{A}_X \to \{0, 1\}^+$ a code for X. We say c is a:

- Uniform Code if $l(x)$ is the same for all $x \in \mathcal{A}_X$
- Var

Types of Codes

Let X be an ensemble and $c : \mathcal{A}_X \to \{0,1\}^+$ a code for X. We say c is a:

- Uniform Code if $l(x)$ is the same for all $x \in \mathcal{A}_X$
- Var

Another i x can be unambiguously determined given $c(x)$.

- Lossless Code if for all $x_1, x_2 \in \mathcal{A}_X$ w $c(x_1) \neq c(x_2)$
- Lossy Code otherwise

Types of Codes

Examples

Examples: Let $\mathcal{A}_X = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$

1. $c(\mathtt{a}) = 00$, $c(\mathtt{b}) = 01$, $c(\mathtt{c}) = 10$, $c(\mathtt{d}) = 11$ is uniform lossless

2. $c(\mathtt{a})$ lossless

3. $c(\mathtt{a}) = 0$, $c(\mathtt{b}) = 0$, $c(\mathtt{c}) = 110$, $c(\mathtt{d}) =$

4. $c(\mathtt{a}) = 00$, $c(\mathtt{b}) = 00$, $c(\mathtt{c}) = 10$, $c(\mathtt{d}) =$

5. $\boxed{c(\mathtt{a}) = -,\ c(\mathtt{b}) = -,\ c(\mathtt{c}) = 10,\ c(\mathtt{d}) = 11 \text{ is uniform lossy}}$

A Note on Lossy Codes & Missing Codewords

When talking about a uniform lossy code c for $\mathcal{A}_X = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ we write

$$c(\mathtt{a}) = 0 \quad c(\mathtt{b}) = 1 \quad c(\mathtt{c}) = -$$

where the symbol − means “no codeword”. This is shorthand for “the receiver

For the pur

$$c(\mathtt{a}) = 0 \quad c(\mathtt{b}) = 1$$

and the sender and receiver agreeing that the code ways be decoded as $\mathtt{b}$

Assigning the outcome $\mathtt{a}_i$ the missing codeword “−” just means “it is not possible to send $\mathtt{a}_i$ reliably”

1. Introduction
   - Overview
   - What is Compression?
   - A Communication Game
   - What's the best we can do?

2. Form
   - En
   - Defining Codes

3. Formalising Compression
   - Reliability vs. Size
   - Key Result: The Source Coding Theorem

Lossless Coding

Example: Colours

Three colour ensemble with $\mathcal{A}_X = \{\texttt{r}, \texttt{g}, \texttt{b}\}$
with $\texttt{r}$ twice as likely as $\texttt{b}$ or $\texttt{g}$

-

Suppos

$$c(\texttt{r}) = 00;\ c(\texttt{g}) = 10;\ \text{and } c(\texttt{b}) = 11$$

For example $c(\texttt{rrgbrbr}) = 00001011001$

On average, we will use $l(\texttt{r})p_{\texttt{r}} + l(\texttt{g})p_{\texttt{g}} + l(\texttt{b})p_{\texttt{b}} = 2$ bits per outcome

- $2N$ bits to code a sequence of N outcomes

Raw Bit Content

Uniform coding gives a crude measure of information: the number of bits required to assign equal length codes to each symbol

Raw Bit Content

If X is an ensemble with outcome set $\mathcal{A}_X$ then its **raw bit content** is

x	c
a	000
b	001
c	010
d	011
e	100
f	101
g	110
h	111

This is a uniform encoding of o

$\mathcal{A}_X = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}, \mathtt{f}, \mathtt{g},$

- Each outcome is enco its
- The probabilities of the outcomes are ignored
- Same as assuming a uniform distribution

For the purposes of compression, the exact codes don't matter – just the number of bits used.

Lossy Coding

Example: Colours

Three colour ensemble with $\mathcal{A}_X = \{\mathtt{r}, \mathtt{g}, \mathtt{b}\}$

- $p_{\mathtt{r}} = 0.5$ and $p_{\mathtt{g}} = p_{\mathtt{b}} = 0.25$.

Exampl

$c(\mathtt{rrrrrrr}) = 0000000$; $c(\mathtt{rrbbrbr}) = 0$

Lossy Coding

Example: Colours

Three colour ensemble with $\mathcal{A}_X = \{\mathtt{r}, \mathtt{g}, \mathtt{b}\}$

- $p_{\mathtt{r}} = 0.5$ and $p_{\mathtt{g}} = p_{\mathtt{b}} = 0.25$.

-

Exampl

$c(\mathtt{rrrrrrr}) = 0000000$; $c(\mathtt{rrbbrbr}) = 0$

What is probability we can reliably code a sequence ?

Given we can code a sequence of length N, how many bits are expected?

What is probability we can reliably code a sequence of N outcomes?

$$P(x_1 \dots x_N \text{ has no } \mathrm{g}) = P(x_1 \neq \mathrm{g}) \dots P(x_N \neq \mathrm{g}) = (1 \quad p_{\mathrm{g}})^N$$

Lossy Coding

Example: Colours

What is probability we can reliably code a sequence of N outcomes?

$$P(x_1 \dots x_N \text{ has no } \mathtt{g}) = P(x_1 \neq \mathtt{g}) \dots P(x_N \neq \mathtt{g}) = (1 \quad p_{\mathtt{g}})^N$$

Given we c cted?

$$\mathbb{E}[I(X_1) + \cdots + I(X_N) | X_1 \neq \mathtt{g}, \dots, X \quad / \quad / \, \mathtt{g}]$$

$$= N \left(I(\mathtt{r}) p_{\mathtt{r}} + I(\mathtt{b}) p_{\mathtt{b}} \right) / (1 - p_{\mathtt{g}}) = N$$

since $I(p_{\mathtt{r}}) = I(p_{\mathtt{b}}) = 1$ and $p_{\mathtt{r}} + p_{\mathtt{b}} = 1 - p_{\mathtt{g}}$.

- c.f. $2N$ bits with lossless code

Lossy Coding

Example: Colours

What is probability we can reliably code a sequence of N outcomes?

$$P(x_1 \ldots x_N \text{ has no } \mathrm{g}) = P(x_1 \neq \mathrm{g}) \ldots P(x_N \neq \mathrm{g}) = (1 \quad p_{\mathrm{g}})^N$$

Given we c … cted?

$$\mathbb{E}[I(X_1) + \cdots + I(X_N) | X_1 \neq \mathrm{g}, \ldots, X \quad / \quad \ldots \quad / \ \mathrm{g}]$$

$$= N\,(I(\mathtt{r})p_{\mathtt{r}} + I(\mathtt{b})p_{\mathtt{b}})/(1 - p_{\mathrm{g}}) = N$$

since $I(p_{\mathtt{r}}) = I(p_{\mathtt{b}}) = 1$ and $p_{\mathtt{r}} + p_{\mathtt{b}} = 1 - p_{\mathrm{g}}$.

- c.f. $2N$ bits with lossless code

Lossy Coding

Example: Colours

What is probability we can reliably code a sequence of N outcomes?

$$P(x_1 \dots x_N \text{ has no } \mathrm{g}) = P(x_1 \neq \mathrm{g}) \dots P(x_N \neq \mathrm{g}) = (1 \quad p_{\mathrm{g}})^N$$

Given we c ... cted?

$$\mathbb{E}[I(X_1) + \dots + I(X_N) | X_1 \neq \mathrm{g}, \dots, X \quad / \quad / \; \mathrm{g}]$$

$$= N\,(I(\mathrm{r})p_{\mathrm{r}} + I(\mathrm{b})p_{\mathrm{b}})/(1 - p_{\mathrm{g}}) = N =$$

since $I(p_{\mathrm{r}}) = I(p_{\mathrm{b}}) = 1$ and $p_{\mathrm{r}} + p_{\mathrm{b}} = 1 - p_{\mathrm{g}}$.

- c.f. $2N$ bits with lossless code

Essential Bit Content

There is an inherent trade off between the number of bits required in a uniform lossy code and the probability of being able to code an outcome

Smallest δ-sufficient subset

Let X be lest
subset of

For small δ, **smallest** collection of **most lik**

Essential Bit Content

There is an inherent trade off between the number of bits required in a uniform lossy code and the probability of being able to code an outcome

Smallest δ-sufficient subset

Let X be lest

subset of

For small δ, **smallest** collection of **most lik**

If we uniformly code elements in S_δ, and ign

- We can code a sequence of length N with probability $(1-\delta)^N$
- If we can code a sequence, its expected length is $N \log_2 |S_\delta|$

Essential Bit Content

Example

Intuitively, construct S_δ by removing elements of $\mathcal{X}$ in ascending order of probability, till we have reached the $1-\delta$ threshold

x	
a	
b	
c	
d	3/16
e	1/64
f	1/64
g	1/64
h	1/64

- $a_i)$

$) \geq 1-\delta$

δ

Essential Bit Content

Example

Intuitively construct S_δ by removing elements of $\mathcal{X}$ in ascending order of probability, till we have reached the $1-\delta$ threshold

x	
a	
b	
c	
d	3/16
e	1/64
f	1/64
g	1/64

- $a_i)$ $) \geq 1-\delta$ δ

  $\delta = 1/64$: $S_\delta = \{\mathtt{a},$

Essential Bit Content

Example

Intuitively construct S_δ by removing elements of $\mathcal{X}$ in ascending order of probability, till we have reached the $1-\delta$ threshold

x	
a	
b	
c	
d	3/16

- $a_i)$

  $) \geq 1 - \delta$

δ

$\delta = 1/64$: $S_\delta = \{$a,

$\delta = 1/16$: $S_\delta = \{$a,

Intuitively, construct S_δ by removing elements of X in ascending order of probability, till we have reached the $1 - \delta$ threshold

x
a

$a_i)$

$) \geq 1 - \delta$

δ

$\delta = 1/64$: $S_\delta = \{$a,

$\delta = 1/16$: $S_\delta = \{$a,

$\delta = 3/4$: $S_\delta = \{$a

Essential Bit Content

Trade off between a probability of δ of not coding an outcome and size of uniform code is captured by the essential bit content

Essential Bit Content

For an ensemble X and $0 \leq \delta \leq 1$, the **essential bit content** of X is

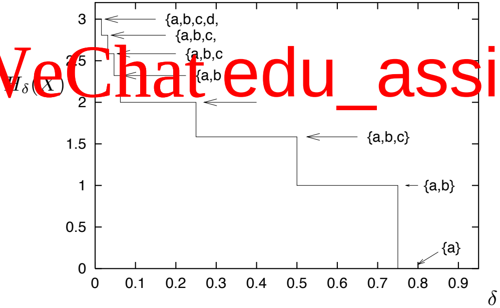
Essential Bit Content

Trade off between a probability of δ of not coding an outcome and size of uniform code is captured by the essential bit content

Essential Bit Content

For an ensemble X and $0 \leq \delta \leq 1$, the **essential bit content** of X is

x	
a	1/4
b	1/4
c	1/4
d	3/16
e	1/64
f	1/64
g	1/64
h	1/64



The Source Coding Theorem for Uniform Codes

(Theorem 4.1 in MacKay)

Our aim next time is to understand this:

The Source Coding Theorem for Uniform Codes

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta <$ $> N_0$

The Source Coding Theorem for Uniform Codes

(Theorem 4.1 in MacKay)

Our aim next time is to understand this:

The Source Coding Theorem for Uniform Codes

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta <$ $> N_0$

What?

- The term $\frac{1}{N} H_\delta(X^N)$ is the average numb uniformly code all but a proportion δ
- Given a tiny probability of error δ, the average bits per symbol can be made as close to H as required.
- Even if we allow a large probability of error we cannot compress more than H bits ber symbol.

Some Intuition for the SCT

- Don't code individual symbols in an ensemble; rather, consider sequences of length N.

- As le
seq

- Thus, we can get by with essentially assigning
each typical sequence

Next time

Recap: typical sets

Formal statement of source coding theorem

Proof of source coding theorem