情報理論第二（3）

情報源符号化とデータ圧縮

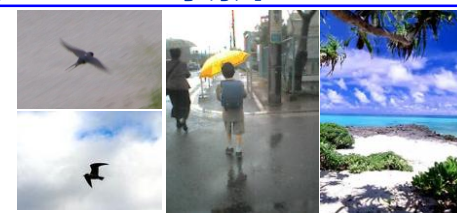
人間コミュニケーション学科
梶本　裕之
kajimoto@hc.uec.ac.jp

レポート回収

授業の流れ（予定変更あり！）

第4週（11/04）　情報源符号化とデータ圧縮
第5週（11/11）　出張のため休講
第6週（11/18）　調布祭のため休講
第7週（11/25）　ハフマン符号とデータ圧縮
第8週（12/ 2）　情報源符号化定理
第9週（12/09）　出張のため休講
第10週（12/16）　マルコフ情報源モデル
第11週（12/23）　休日
第12週（ 1/ 6）　通信路のモデル化
第13週（ 1/13）　センター試験準備のため休講　←変更点
第14週（ 1/20）　誤り検出・誤り訂正符号
第15週（ 1/27）　線形符号
第16週（ 2/ 3）　ハミング符号，秘密鍵暗号
第17週（ 2/10）　公開鍵暗号　←変更点

前回：二つの事象系



- 事象系　X = ツバメが{低空飛行する，しない}
- 事象系　Y =雨が{降る，降らない}
- 結合エントロピーH(X,Y)
- 条件付エントロピーH(X|Y)
- 相互情報量 I(X:Y)= I(Y:X)=H(Y)−H(Y|X)
  二つの事象系の「結びつきの強さ」を表す量

前回の小レポート(1)

$$I(Y:X) = H(Y) - H(Y \mid X)$$
$$= H(Y) + H(X) - H(YX)$$

上記の等式が成り立つことを，それぞれのエントロピーの定義に立ち返って計算し，確かめよ．

前回の小レポート(1)：回答例

$$I(Y:X) = H(Y) - H(Y \mid X)$$
$$= H(Y) + H(X) - H(YX)$$

$$-H(Y \mid X) = \Sigma_i \Sigma_j [\, p(x_i, y_j) \log p(y_j \mid x_i) \,]$$
$$= \Sigma_i \Sigma_j [\, p(x_i, y_j) \log \{p(x_i, y_j)/p(x_i)\} \,]$$
$$= \Sigma_i \Sigma_j p(x_i, y_j) \log p(x_i, y_j) - \Sigma_i p(x_i) \log p(x_i)$$
$$= -H(XY) + H(X)$$

前回の小レポート(2)：回答例

- 相互情報量 $I(X;Y)$ がとりうる値の範囲は $0 \leqq I(X;Y) \leqq \min\{H(X), H(Y)\}$ であることが知られている．この最小値・最大値がそれぞれどのような場合に与えられるかを説明せよ．

7

前回の小レポート(2)：回答例

$I(X;Y) = H(Y) - H(Y \mid X)$
$= H(X) - H(X \mid Y)$

- XとYが完全に独立のとき
  $H(Y \mid X) = H(Y)$，$H(X \mid Y) = H(X)$ となるので $I(X;Y) = 0$

ダメなツバメ

	$X_{低}$	$X_{高}$
$Y_{雨}$	0.2	0.2
$Y_{晴}$	0.3	0.3

- YがXに従属であるとき $H(Y \mid X) = 0$ となるので
  $I(X;Y) = H(Y)$（このとき常に $H(X) \geqq H(Y)$）

優秀なツバメ

	$X_{低}$	$X_{高}$
$Y_{雨}$	0.4	0
$Y_{晴}$	0	0.6

8

前回の小レポート(3)

- 身近にある，互いに影響を及ぼしあっていると思われる現象を2つ選び，それぞれ確率的事象系として考え，その間の相互情報量を計算し，事象系の間にある関連性の強さを評価せよ．

  - 例：苗字の画数と名前の画数　など

略

9

情報源符号化

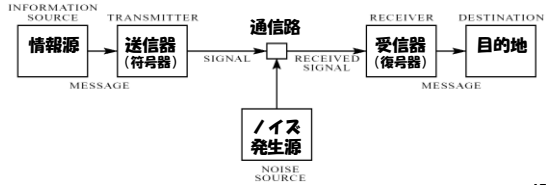
授業の流れ

- 前回まで：
  事象系が持つ「情報の量」を
  定量的に測る方法を見てきた

- 今週から：
  事象系が生み出す情報を
  効率よく記号で表す（符号化）
  方法を考える

11

シャノンの情報通信のモデル



情報源符号化

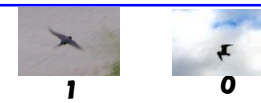
- 確率的と見なしうる事象を次々に生み出すシステム（情報源）の振る舞いを，何らかの記号群を用いて表現する
  - 時間的に生み出すもの：音声など
  - 空間的に生み出すもの：静止画像など
  - 時間空間的に生み出すもの：動画など

13

例

確率的な事象系（情報源）

記号

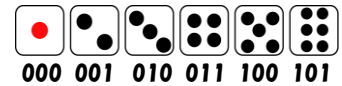


情報源の振る舞い

記号表現

確率的な事象系（情報源）

記号



情報源の振る舞い

記号表現

14

用語の定義

- 符号語：ある事象に対応させた記号(列)
- 符号長：ある符号語を構成する記号の数（符号語の長さ）
- 符号　：対応する事象と符号語のペアの集合（両者間の変換規則全体）
- 符号化：事象→符号語への変換
- 復号化：符号語→事象への変換

15

例：ASCII 符号

- 事象：英語のアルファベットなど
- 符号語に用いられる記号＝{0,1}
- “a”の符号語＝“1100001”
- 符号長＝7ビット
- 符号：右の表全体
- 符号化：“a”→“1100001”の操作
- 復号化：“1100001”→“a”の操作

上位3ビット→ ↓下位4ビット	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAC	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF/NL	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	¥	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

16

固定長符号と可変長符号

- 固定長符号：
  全ての事象に対して符号語の長さが常に等しいような符号
- 可変長符号：
  表現する事象に応じて符号語の長さが変わるような符号

17

固定長符号の例

- ASCII（常に1文字7ビット）
- 普通の英文テキストファイル（常に1文字8ビット＝1バイト）
- PNM [PPM, PGM, PBM] フォーマットによる画像データ（1ピクセルごとに色情報を数値化して並べた形式）

18

可変長符号の例

A ・－　I ・・　Q ・－・－　Y －・－－
B －・・・　J ・－－－　R ・－・　Z －－・・
C －・－・　K －・－　S ・・・　. 終点
D －・・　L ・－・・　T －　, 小読点
E ・　M －－　U ・・－　: 重点又は除去の記号
F ・・－・　N －・　V ・・・－　? 問符
G －－・　O －－－　W ・－－
H ・・・・　P ・－－・　X －・・－

- モールス信号
  "SOS" → ・・・　－－－　・・・

- ほとんどの画像や音声ファイル

19

例題

- n個の文字（事象）を2進数表記による固定長符号で符号化することを考える．必要な符号長を求めよ．
- また，符号化の効率（使用可能な符号語数に対する実際に使用されている符号語数の割合）は，nがどのような場合に最良・最悪となるか．

20

例題

n =

- 2: →必要な符号長
- 3: →必要な符号長
- 4: →必要な符号長
- 5: →必要な符号長

結局，必要な符号帳は

最良：

最悪：

21

今日の世界：「あいうえお世界」

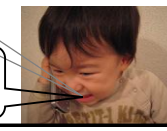
A,I,U,E,Oの５文字だけの世界

あーもしもし私だが例のプロジェクトはどうした？

a oioi aai aa ei o uoeuo a ouia?

申し訳ございません．現在鋭意努力しておりまして，明日までには必ず．

Ouiae oaiaeu
Euai eii oouie oiaie
Au ae ia aaau.

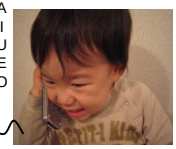


あいうえお世界の固定長符号の例

A I U E O → 符号化

A	000
I	001
U	010
E	011
O	100

→ 複号化 → A I U E O



A oioi aai aa ei o uoeuo a ouia?
000,100,001,100,001,000,000,001,000,000,011,001,100,010,100,011,010,100,000,100,010,001,000

23文字⇒23×３＝69bit

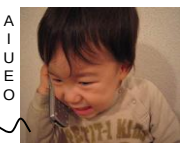
23

あいうえお世界の可変長符号の例

A I U E O → 符号化

A	00
I	10
U	110
E	111
O	01

→ 複号化 → A I U E O



A oioi aai aa ei o uoeuo a ouia?
00,01,10,01,10,00,00,10,00,00,111,10,01,110,01,111,110,01,00,01,110,10,00

23文字⇒23×３＝51bit

24

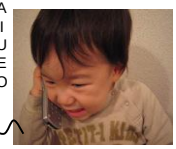
問題（今日の話）

A I U E O → 符号化

A	000
I	001
U	010
E	011
O	100

A	00
I	10
U	110
E	111
O	01

復号化 → A I U E O



固定長：69bit⇒可変長：51bit

約74%の情報圧縮に成功!!

- ちゃんと元に戻せるの？
- どこまで圧縮できるの？

25

今日のお題

- 符号木とプレフィックス符号
- Kraft の不等式
- データ圧縮とその限界

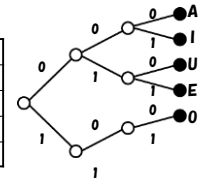
26

符号木とプレフィックス符号

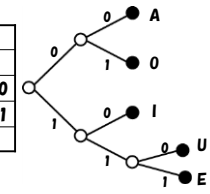
符号木

- 符号語から元の文字を検索（復号）するのに用いられる木

A	000
I	001
U	010
E	011
O	100



A	00
I	10
U	110
E	111
O	01



黒いノードは符号語を表す

28

例題

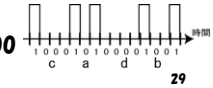
- 符号木を作成せよ

(1)ツバメが低く飛ぶ→0, 高く飛ぶ→1

(2)サイコロの出る目を2進符号化

(3)あいうえお世界の符号化
A→0, I→100, U→101, E→1, O →11

(4)パルス幅符号化
a→10, b→100, c→1000, d→10000



29

例題

(1)ツバメが低く飛ぶ→0, 高く飛ぶ→1

(2)サイコロの出る目を2進符号化

30

例題

(3)あいうえお世界の符号化
A→0, I→100, U→101, E→1, O →11

(4)パルス幅符号化
a→10, b→100, c→1000, d→10000

31

符号木と復号化

- 入力に応じて木を探索する
  - 末端でも符号語でもないノード：入力に従って次のノードに移動する
  - 末端の符号語ノード：対応する文字を出力して根（ルートノード）に戻る
  - 末端でない符号語ノード：2つの状態（対応する文字を出力して根に戻る状態と，文字を出力せずその先に進む状態）に分岐して並行に探索を続ける
  - 末端で符号語でないノードに来たとき，エッジのない方向への入力があったとき，及び符号語に至らない段階で入力が無くなったときはその探索を停止する（その状態が消滅する）

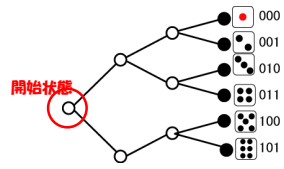
32

例1（等長符号）

- 符号化された情報を復号化する動作を，符号木上でシミュレートしてみる．

001100

情報　2-5
符号化　001100
復号化　2-5

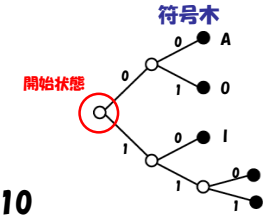


33

例2（可変長符号-1）

- 可変長符号では，もはや符号の「区切り」がどこか分からない

A	00
I	10
U	110
E	111
O	01



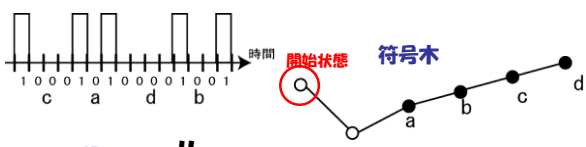
情報　OAOU
符号化　0100 01110
復号化　OAOU

復号成功！！

34

例3（可変長符号-2）

- パルス幅符号化
  a→10, b→100, c→1000, d→10000



情報　cadb
符号化　1000101000 0100
復号化　cadb

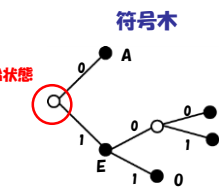
復号成功！！

35

例4

- ダメな場合

A	0
I	100
U	101
E	1
O	11



入力　OAOU
符号化　11011101
復号化　EEAEEEAE

復号失敗！！
（“区切り文字” が無いことに注意！）

36

復号可能性

- ある符号によって符号化された任意の（有限の）記号列が，常に一意に復号されるならば，その符号は**復号可能である**

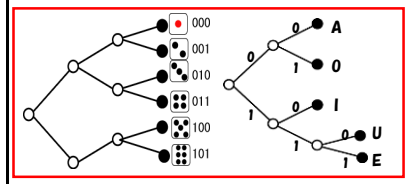
＝「どちらともとれる」ような状況が生じない

符号は復号可能でなければ意味がない

37

プレフィックス符号

符号語のノードが**終端ノード**に限られている符号



木の途中に符号語ノードがない→ 一意に復号可能

Prefix=接頭語：「どの符号も他の符号の接頭語になっていない」という意味から．

実はプレフィックス符号によって「最も効率の良い符号化」が実現できることが分かっている

38

電話番号はプレフィックス符号

- 電話番号は可変長．
- ある番号をかけている最中に，別の番号にかかってしまったら問題．

Aさん：080-1234-5678
Bさん：080-1234-567

Aさんにかけられない？！

39

Kraft の不等式

Kraft の不等式

- 事象数：n
- 各事象に対する符号長：L_i $(i=1\sim n)$

プレフィックス符号
（あるいは復号可能な符号）
が存在する必要十分条件は

$$\Sigma_{i=1\sim n} 2^{-L_i} \leqq 1$$

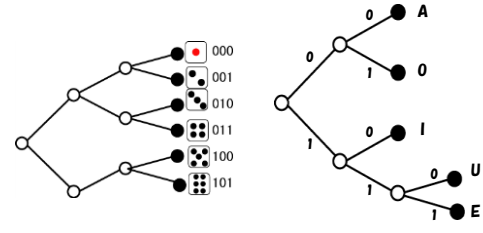
AIUEO：n=5

	符号	符号長
A	00	2
I	10	2
U	110	3
E	111	3
O	01	2

41

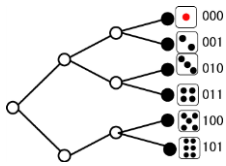
例題

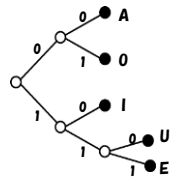
- 下記の例について，Kraft の不等式 $\Sigma_i 2^{-L_i} \leqq 1$ が成立することを確かめよ．



42

例題

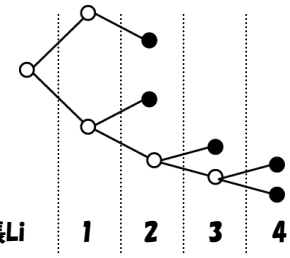




43

Kraftの不等式の意味(1)

Kraft の不等式　$\Sigma_i 2^{-Li} \leqq 1$

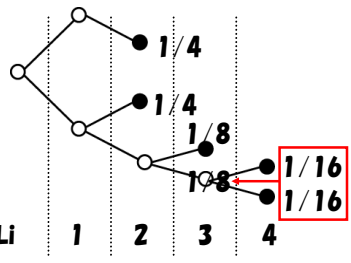


符号長Li	1	2	3	4
2^{-Li}	1/2	1/4	1/8	1/16

44

Kraftの不等式の意味(2)

Kraft の不等式　$\Sigma_i 2^{-Li} \leqq 1$



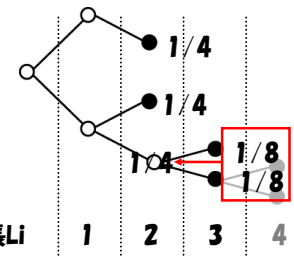
符号長Li　1　2　3　4

右から順に足していく

45

Kraftの不等式の意味(3)

Kraft の不等式　$\Sigma_i 2^{-Li} \leqq 1$



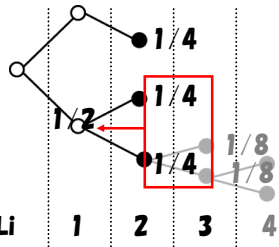
符号長Li　1　2　3　4

右から順に足していく

46

Kraftの不等式の意味(4)

Kraft の不等式　$\Sigma_i 2^{-Li} \leqq 1$



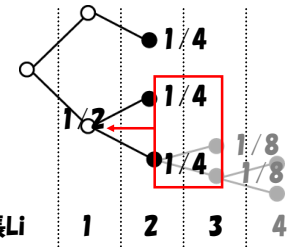
符号長Li　1　2　3　4

右から順に足していく

47

Kraftの不等式の意味(5)

Kraft の不等式　$\Sigma_i 2^{-Li} \leqq 1$



符号長Li　1　2　3　4

右から順に足していく

48

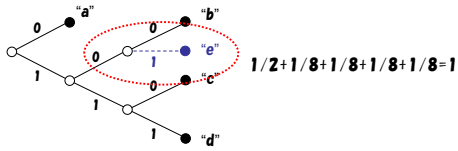
Kraft の不等式と完全な符号

Kraftの不等式において等号が成立する

⇕

符号が「完全である」

（符号木に未使用の終端ノードがない）



データ圧縮とその限界

符号化によるデータの圧縮

符号化手法を工夫して，データ
（情報源が生成する具体的な事象列）
をなるべく短く表すことを考える

可変長符号の平均符号長

- 情報源における各事象の発生頻度が既知であれば1事象あたりの符号長の期待値（平均符号長）を計算することができる

$$L = \Sigma_i \, p_i \, L_i$$

（p_i：事象 i の出現確率）

例題

- 平均符号長を求めよ．

A	00
I	10
U	110
E	111
O	01

A oioi aai aa ei o uoeuo a ouia?
00,01,10,01,10,00,00,10,00,00,111,10,01,110,01,111,110,01,00,01,110,10,00

例題

- 平均符号長を求めよ．

A oioi aai aa ei o uoeuo a ouia?
00,01,10,01,10,00,00,10,00,00,111,10,01,110,01,111,110,01,00,01,110,10,00

- 全長：
- 語数：
-

平均符号長とデータ圧縮

$$L = \Sigma_i p_i L_i$$

(p_i：事象 i の出現確率)

データ圧縮とは：

情報源の統計的な性質 p_i を調べ，
それに応じて平均符号長 L を最小化する
「良い符号」を構築し，
それを用いてデータを符号化すること

55

参考：可逆圧縮と非可逆圧縮

- ここで議論しているのは「可逆圧縮」
  （データを完全に復元することが可能な圧縮方法）
- 音声や画像の圧縮では，人間の知覚における限界を利用して品質を余り落とさないままデータを減らす「非可逆圧縮」も用いられている（例：JPEG, MPEG, MP3）

56

圧縮効率をよくするには？

$$L = \Sigma_i p_i L_i$$

(p_i：事象 i の出現確率)

基本方針：

頻出する事象には短い符号語を
めったに生じない事象には長い符号語を
割り当てる

（具体的なアルゴリズムは次回やります）

57

平均符号長はどこまで縮められるか

- ある情報源に対する復号可能な符号化の方法の中で平均符号長を最小にするものがあると仮定すると

$$L_{min} = \min_{\text{復号可能な符号}} \Sigma_i p_i L_i$$

$$L_{min} = \min_{\text{Kraftの不等式を満たす}\{L_i\}} \Sigma_i p_i L_i$$

58

平均符号長はどこまで縮められるか

$$L_{min} = \min_{\text{Kraftの不等式を満たす}\{L_i\}} \Sigma_i p_i L_i$$

L_i が整数に限られている最適化問題であるため，一般解を得るのは容易でない

⬇

L_i を連続変数 x_i で置き換えれば，Kraft の式の等号が成り立つ（符号が完全である→無駄が全く無い）ような $\{x_i\}$ によって L_{min} の下界が与えられる

59

平均符号長はどこまで縮められるか

L_i を連続変数 x_i で置き換えれば，Kraft の式の等号が成り立つ（符号が完全である→無駄が全く無い）ような $\{x_i\}$ によって L_{min} の下界が与えられる

最小化する量：　$\Sigma_i p_i x_i$
制約条件：　$\Sigma_i 2^{-x_i} = 1$

（条件付き極値問題→ラグランジュの未定乗数法を使う）

60

平均符号長はどこまで縮められるか

- 次のような n+1 変数関数を用意する.

$$g(x_1, x_2, \cdots, x_n, x_{n+1}) = \underbrace{\sum_{i=1}^{n} p_i x_i}_{\text{最小化する量}} + x_{n+1}\underbrace{(\sum_{i=1}^{n} 2^{-x_i} - 1)}_{\text{制約条件}}$$

- 全ての i について極値条件
  $\partial g / \partial x_i = 0$ が成立する条件を考える

平均符号長はどこまで縮められるか

$$g(x_1, x_2, \cdots, x_n, x_{n+1}) = \sum_{i=1}^{n} p_i x_i + x_{n+1}(\sum_{i=1}^{n} 2^{-x_i} - 1)$$

(i = 1~n) $\dfrac{\partial g}{\partial x_i} = p_i - x_{n+1} 2^{-x_i} \ln 2 = 0$ (i = n+1) $\dfrac{\partial g}{\partial x_{n+1}} = \sum_{i=1}^{n} 2^{-x_i} - 1 = 0$

$$2^{-x_i} = \frac{p_i}{x_{n+1} \ln 2} \longrightarrow \sum_{i=1}^{n} 2^{-x_i} - 1 = \sum_{i=1}^{n} \frac{p_i}{x_{n+1} \ln 2} - 1$$

$$= \frac{1}{x_{n+1} \ln 2} \sum_{i=1}^{n} p_i - 1$$

$$2^{-x_i} = p_i \qquad = \frac{1}{x_{n+1} \ln 2} - 1 = 0$$

$$x_{n+1} = \frac{1}{\ln 2}$$

$$x_i = -\log p_i$$

最小平均符号長の下界

$$x_i = -\log p_i$$

$L_{min} \geqq \min \Sigma_i p_i x_i$

$= - \Sigma_i p_i \log p_i$ 情報源が持つエントロピー

結論： 情報源符号化において
その平均符号長を
情報源のエントロピーよりも
小さくすることはできない

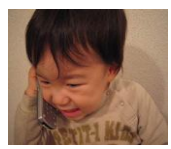
言い換えれば. . .

エントロピーとは,
その事象系が生みだしている
正味の情報量.
それよりも小さくデータ
圧縮することはできない.

（データ圧縮とはつまるところ
無駄を省いているに過ぎない）

小レポート

- あいうえお世界ではなく，「あいうえおっん」世界を考える．（文字数7）
  100文字程度の日常的な会話例文を作成し，そこで事象系全体がもつ情報エントロピーを定量的に評価し，平均符号長がそれになるべく近くなるように符号を具体的に設計してみよ.



次回は11/25