今日の講義のねらい (1)

物理専攻だとシャノン流の情報理論を
ぜんぜん習っていない (かもしれない)

やはりいちどは聞いておくべきもの

■シャノン理論
情報圧縮の理論 (雑音なし) 今回こちら
誤り訂正の理論 (雑音あり)

情報圧縮

身近な話題になってきている

テキスト lha,gzip,zip,bzip2 (可逆)
画像 jpeg(非可逆) png,gif (可逆)

ここでは「可逆圧縮」(100%戻る)
に限って論じる

今日の講義のねらい (2)

@ むかしの考え方
文字単位の圧縮→ブロック単位
@ 新しい考え方
テキスト全体→モデル化→実効的に部分に分割

予測 = 圧縮という見方

従来の展開にほぼ従いながら, この2つをたえず意識

今日の講義のねらい (3)

情報理論入門の多くでは
「シンボルの出現確率は既知」
「適当に数えればわかる」
としている

(後半)
確率未知 → 統計科学との接点
MDL原理

参考書

@ やさしい本
大石進一 例にもとづく情報理論入門 講談社
甘利俊一 情報理論 ダイヤモンド社 (版切れ)
@ 薄いが本格的な本
情報源符号化—無歪みデータ圧縮 培風館

@ 後半 (MDLなど) について 上の本の6章にもあり
統計科学のフロンティア3 モデル選択 岩波書店
(第2部 伊藤秀一 確率的複雑さとMDL原理)

@ 最新の動向含む専門的なレビュー (IBIS2001,Webにあり)
ユニバーサルデータ圧縮アルゴリズムの変遷
—基礎から最新手法まで— 山本博資

I　情報圧縮はやわかり

夢の圧縮法？

すべてのファイルを1/100のサイズに圧縮します

詐欺

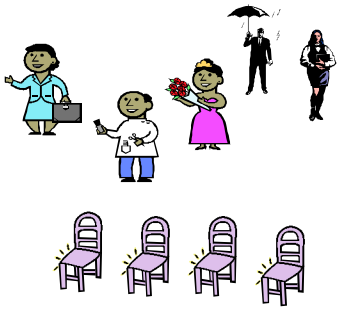
長さ1000ビットのファイル　2^{1000} 個
長さ10　ビットのファイル　2^{10} 個
長さ999　ビットのファイル　2^{999} 個

N個のものをN-1個に入れたら．．

必ずどれか重複する
⇒ 可逆圧縮ではありえない

「鳥の巣箱論法」
椅子のほうが人より少なければ
誰か座れない人が出る

かならず人のほうががあまる



「‥以下の長さ」でもだめ

●●●●●
●●●●　●●●●
●●●　●●●
●●　●●
●　●

なぜ可逆圧縮できるか

原理
出現確率の低い対象には長いコードを
出現確率の高い対象には短いコードを
割り当てればよい

全部短くする　→　区別ができなくなるからだめ

古典的な例： 英字の頻度

e	11.4	12.3
t	8.2	9.1
a	8.4	8.1

多いほう(%)

j	0.21	0.2
q	0.08	0.1
z	0.08	0.07

少ないほう(%)

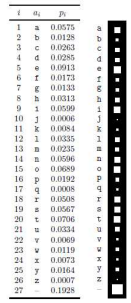


Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

David J.C. MacKay
Information Theory, Inference, and Learning Algorithms
より抜粋

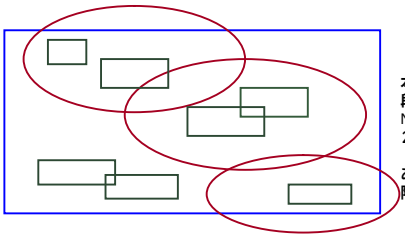
モールス符号

欧　　文

文字		文字	
・－	A	－・	N
－・・・	B	－－－	O
－・－・	C	・－－・	P
－・・	D	－－・－	Q
・	E	・－・	R
・・－・	F	・・・	S
－－・	G	－	T
・・・・	H	・・－	U
・・	I	・・・－	V
・－－－	J	・－－	W
－・－	K	－・・－	X
・－・・	L	－・－－	Y
－－	M	－－・・	Z

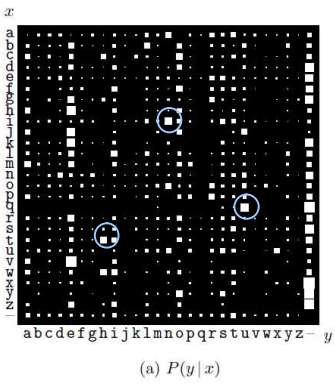
文字の相関

THE とかHEとかいう言葉がたくさんある
⇒ Hのあとは**E**が多いはず



本，記事や章，
段落，文，単語，
N文字，. . . ，3文字，
2文字

あらゆるレベルで
階層的に相関構造



(a) $P(y|x)$

David J.C. MacKay
Information Theory, Inference, and Learning Algorithms
より抜粋

非独立性の表現

これらをとりこむことでより圧縮できる

低レベルの相関構造の表現の
ひとつの方法はブロックの確率
b a a a b a a a a a b a b a a b b a b a
ba aa ba aa aa ba ba ab ba ba
baa aba aaa aba baa bba ba

条件つき確率で表現

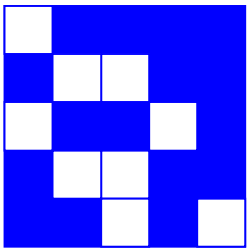
少し違う方法としてマルコフ連鎖や
条件つき確率を使うこともできる

$P(x_i|x_{i-1})$ マルコフ連鎖

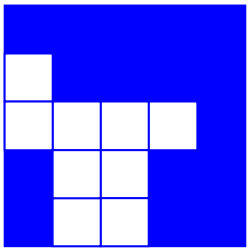
$P(x_i|x_{i-1}, x_{i-2}, \cdots, x_{i-m})$ m次

$P(x_i|x_{i-1} \cdots\cdots x_0)$ 過去の全部

画像



画像



同じ色が固まった画像なら
境界線を符号化したほうが
よいかもしれない.

境界線は珍しい
⇒ 確率が小さい
⇒ 符号が長い

画像:条件つき確率

$P(p_0|p_1, p_2, \ldots, p_{17})$

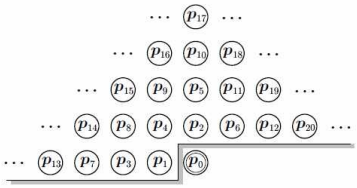


図 1 画素配置

「予測」と符号化

よくおこる事象 → 短い符号
おこりにくい事象 → 長い符号

別の見方
予測のつくことは予測してもらう
送り手と受け手が同じ予測器を持つ
⇒ 予測不能のときに送る

確率を計算 ⇒ 予測する というふうに考える

II. 理想符号長と情報量

最短平均符号長の導出（発見的）

2つの事象が独立なら　定義から

$$P(x_1, x_2) = P(x_1)P(x_2)$$

2つの事象が独立なら　たぶん　記述長は和

$$l(x_1, x_2) = l(x_1) + l(x_2)$$

関数方程式

$P(x) = \widetilde{P}(l(x))$ と書けるとする

$$\widetilde{P}(l_1)\widetilde{P}(l_2) = \widetilde{P}(l_1 + l_2)$$

$$\log \widetilde{P}(l_1) + \log \widetilde{P}(l_2) = \log \widetilde{P}(l_1 + l_2)$$

$\log \widetilde{P}(l) = -a \times l$　　$l(x) = -\frac{1}{a}\log(P(x)) + c$

定数　a c

定数

$$l(x) = -\frac{1}{a}\log(P(x)) + c$$

$x = 0, 1, ..., 2^L - 1,\ P(x) = \frac{1}{2^L}$　　$l = L$

$P(\text{“0”}) = 1$　　$l = 0$

$l(x) = -\log_2(P(x))$　　2文字（0と1）でコードした符号長の場合

以下log と書いたら2が底と約束する

理想符号長

符号の長さを確率の対数にマイナスをつけたものに比例して取るのが自然

$$l(x) = -\log P(x)$$

確率のけた数

整数にならないじゃん！⇒　あとで考える

シャノン情報量

理想的な符号の符号長の期待値

$$-\sum_x P(x)\log P(x)$$

シャノン情報量とよばれる

マルコフ連鎖の場合

$$P(x_1, x_2, \cdots, x_{n-1}, x_n) =$$
$$P(x_n|x_{n-1})P(x_{n-2}|x_{n-3})\cdots P(x_3|x_2)P(x_2|x_1)P(x_1)$$

$$\log P(x_1, x_2, \cdots, x_{n-1}, x_n) =$$
$$\sum_{i=1}^{n} \log P(x_i|x_{i-1}) + \log P(x_1)$$

この値の大小で符号の長さを決めればよい

「予測」という観点から

$P(x)$ を $Q(x)$ と思っているとどれだけ損か

$$\sum_x P(x)\{-\log P(x) - (-\log Q(x))\}$$

$$= -\sum_x P(x)\log\frac{P(x)}{Q(x)} \leq 0$$

↑次の頁で示す

$$-\sum_x P(x)\log P(x) \leq -\sum_x P(x)\log Q(x)$$

不等式の証明

$$\log_e x \leq x-1 \qquad \log_2 x \leq (x-1)\log_2 e$$

$$P(x)\log\frac{P(x)}{Q(x)} =$$

$$-P(x)\log\frac{Q(x)}{P(x)} \leq -P(x)\frac{Q(x)}{P(x)} + P(x) = -Q(x) + P(x)$$

$$-\sum_x P(x)\log\frac{P(x)}{Q(x)} \leq \sum_x Q(x) - \sum_x P(x) = 0$$

カルバック情報量

KL-divergence

$$D(P||Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)} \geq 0$$

一種の分布間の距離

ただし一般にはD(P||Q)≠D(Q||P)

ギブス分布（統計力学）との比較

2つの事象が独立なら　定義から

$$P(x_1, x_2) = P(x_1)P(x_2)$$

2つの事象が独立（？）なら　エネルギーは和　（？）

$$E(x_1, x_2) = E(x_1) + E(x_2)$$

関数方程式

$P(x) = \widetilde{P}(E(x))$　と書けるとする

$$\widetilde{P}(E_1)\widetilde{P}(E_2) = \widetilde{P}(E_1 + E_2)$$

$$\log\widetilde{P}(E_1) + \log\widetilde{P}(E_2) = \log\widetilde{P}(E_1 + E_2)$$

$$\log\widetilde{P}(E) = -\beta\cdot E \quad \beta \text{ 定数}$$

$$P(x) \propto \exp(-\beta E(x)) \quad \text{カノニカル分布}$$

関数方程式（符号の場合）

$P(x) = \widetilde{P}(l(x))$　と書けるとする

$$\widetilde{P}(l_1)\widetilde{P}(l_2) = \widetilde{P}(l_1 + l_2)$$

$$\log\widetilde{P}(l_1) + \log\widetilde{P}(l_2) = \log\widetilde{P}(l_1 + l_2)$$

$\log\widetilde{P}(l) = -a\times l$　　$l(x) = -\frac{1}{a}\log(P(x)) + c$

定数　a　c　　$P(x) \propto \exp(-al(x))$

本当はぜんぜん違う

熱平衡統計力学
- ミクロ　古典力学（リウビルの定理）
- 量子力学
- マクロ　熱力学（を介した経験事実）

情報理論
- 組み合わせの数についての数学的な事実
- →（この部分は）数学的に証明できる

「確率」の基本

それ以前のレベルでのみ，
似ているといえる

独立なら確率は　積
関係なければ和になる量

確率論は積と和のなすドラマである

III. 情報源符号化定理の証明

情報源符号化定理（シャノン）

$$-\sum_{x} P(x)\log P(x)$$

今までの議論は「予想」
このあときちんと示す

@ 平均符号長の下限
@ いくらでも近い符号化が実現可能

原点に戻る

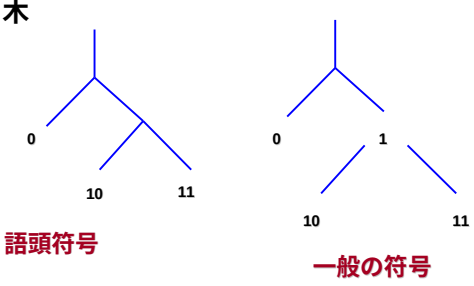
N個のものをN-1個に入れたら．．
必ずどれか重複する
⇒ 可逆圧縮ではありえない

情報理論の数理の要点

「鳥の巣箱論法」
椅子のほうが人より少なければ
誰か座れない人が出る

符号の木

符号の木



語頭符号　　一般の符号

符号を短くする限界 (1)

符号語の長さを $l(x_1), l(x_2), \cdots l(x_n)$

長さ l のものは最大 2^l 個

任意の符号に対して，和 $\sum 2^{-l(x)}$ を作ると

$$\sum_x 2^{-l(x)} \leq \sum_{l=1}^{N} \{2^{-l} \cdot 2^l\} = N$$

すべての符号語についての和

符号語の長さの上限

区切りの問題

欧文

文字		文字	
・－	A	－・	N
－・・・	B	－－－	O
－・－・	C	・－－・	P
－・・	D	－－・－	Q
・	E	・－・	R
・・－・	F	・・・	S
－－・	G	－	T
・・・・	H	・・－	U
・・	I	・・・－	V
・－－－	J	・－－	W
－・－	K	－・・－	X
・－・・	L	－・－－	Y
－－	M	－－・・	Z

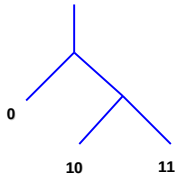
モールス符号は区切りが必要（長く空ける）

分節可能な符号

区切り記号を別に用意しなくても
よい記号のことを分節可能という

語頭符号なら分節可能
(逆は不成立)

0 1 0 0 1 1 0



注意

「分節可能」は「一意複号可能」ともいうが
「多対1にならない」という意味ではない
あくまでも「区切り」の問題
「多対1にならない」符号のことは
「正則符号」という　(⇔非正則)

大きなかたまりで符号化するなら
→分節可能でない正則符号も実用可能

符号を短くする限界 (2)

$$\sum_x 2^{-l(x)} \leq N$$

分節可能ならより強く以下がいえる

$$\sum_x 2^{-l(x)} \leq 1$$

Nによらない！

クラフト・マクミランの不等式

クラフトの不等式

$$\sum_x 2^{-l(x)} \leq 1$$

(a) 分節可能な符号　→クラフトの不等式をみたす

(b) クラフトの不等式をみたす　→符号が構成可能

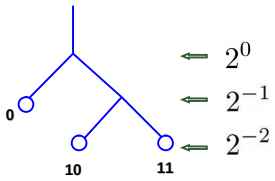
以下で証明する　→　これらから情報源符号化定理が出る

上限： 語頭符号の場合

語頭符号ならほぼ自明

体積1の水を流し込む

$$\sum_x 2^{-l(x)} \le 1$$



語頭符号の木

一般の場合：上限

(a) 分節可能な符号 →クラフトの不等式をみたす

a,bの2文字を符号化したとする
aa,ab,ba,bb 4文字
aaa,aab,aba,abb,baa,bab,bba,bbb 8文字
の符号が作れる
→ 「分節可能」なので区切り不要
単に符号語をくっつければよい

そこで‥

$\sum_x 2^{-l(x)}$ に相当する量は

$$\left(\sum_x 2^{-l(x)}\right)^2 = \left(\sum_x 2^{-l(x)}\right)\left(\sum_{x'} 2^{-l(x')}\right)$$

$$\left(\sum_x 2^{-l(x)}\right)^m$$

すると．．

$$\left(\sum_x 2^{-l(x)}\right)^m \le N$$

$$\sum_x 2^{-l(x)} \le N^{1/m}$$

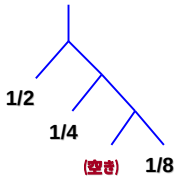
$$\sum_x 2^{-l(x)} \le \lim_{m\to\infty} N^{1/m} = 1$$

証明終

具体的に構成できること

(b) クラフトの不等式をみたす $\{l(x)\}$→符号が構成可能

語頭符号の範囲で
逐次的に構成できる



$$(2^{-l_1}, 2^{-l_2}, 2^{-l_3}) = (1/2, 1/4, 1/8)$$

確率がでかい順（長さが短い順）につっこむのがコツ

クラフトの不等式

$$\sum_x 2^{-l(x)} \le 1$$

(a) 分節可能な符号 →クラフトの不等式をみたす

(b) クラフトの不等式をみたす →符号が構成可能

(a),(b) 証明完了 → これから情報源符号化定理が出る

情報源符号化定理（シャノン）

$$-\sum_x P(x)\log P(x)$$

(i) 平均符号長の下限

(ii) いくらでも近い符号化が実現可能

(i) 符号長の下限

(a) 分節可能な符号→クラフトの不等式をみたす

$$\sum_x 2^{-l(x)} \le 1$$

$$\tilde{Q}(x) = 2^{-l(x)}$$

$$\sum_x \tilde{Q}(x) \le 1$$

「確率もどき」になっている
（劣確率）

確率もどきでも・・

$P(x)$ を $\tilde{Q}(x)$ と思っているとどれだけ損か

$$\sum_x P(x)\left\{-\log P(x) - (-\log \tilde{Q}(x))\right\}$$

$$= -\sum_x P(x)\log\frac{P(x)}{\tilde{Q}(x)} \le 0$$

$$-\sum_x P(x)\log P(x) \le -\sum_x P(x)\log \tilde{Q}(x) = \sum_x P(x)l(x)$$

劣確率の場合の不等式の証明

$$\log_e x \le x-1 \qquad \log_2 x \le (x-1)\log_2 e$$

$$P(x)\log\frac{P(x)}{\tilde{Q}(x)} =$$

$$-P(x)\log\frac{\tilde{Q}(x)}{P(x)} \le -P(x)\frac{\tilde{Q}(x)}{P(x)} + P(x) = -\tilde{Q}(x) + P(x)$$

$$-\sum_x P(x)\log\frac{P(x)}{\tilde{Q}(x)} \le \sum_x \tilde{Q}(x) - \sum_x P(x) \le 0$$

(ii) 理想符号長の実現

$$l(x) = -\log P(x)$$

整数にならないのが問題 → とりあえず丸める

$$l(x) = [-\log P(x)] + 1 \ge \log P(x)$$

$$\sum_x 2^{-l(x)} \le 1$$ 満たす

とりあえず，誤差1以内

(b) クラフトの不等式をみたす → 符号が構成可能

$$\sum_x l(x)P(x) \le [-\sum_x P(x)\log P(x)] + \sum_x P(x)$$

$$\sum_x l(x)P(x) \le \sum_x P(x)\log P(x) + 1$$

ブロック符号化で半端を減らす

こんどはさっきと違って「ブロックを作ってから符号化」

a,bの2文字を符号化するかわりに
aa,ab,ba,bb 4文字
aaa,aab,aba,abb,baa,bab,bba,bbb 8文字
‥ m文字をまとめて符号化する

確率の計算ではシンボルは独立とみなす

ブロック符号化と情報量

$$-\sum_{x_1,x_2} P(x_1)P(x_2)\log P(x_1)P(x_2)+1$$

$$=-\sum_{x_1} P(x_1)\log P(x_1)-\sum_{x_2} P(x_2)\log P(x_2)+1$$

$$=-2\sum_{x} P(x)\log P(x)+1$$

おつりはいつも1

$$\frac{1}{m}\sum_x P(x)l(x)=\frac{1}{m}\left\{m\sum_x P(x)\log P(x)+1\right\}$$

$m\to\infty$

「鳥の巣箱論法」
椅子のほうが人より少なければ
誰か座れない人が出る

いままでの話

1文字づつの符号化を念頭においていたが
実際にはxがなんであっても成り立つ

理論的には！
実際はいろいろ問題点がある ⇒以下で検討

X 単語，パラグラフ，文書全体
時系列全体，画像全体‥

問題その1：クラフト不等式

一意解読可能＝区切り不要 に限定
「1章分まるまる符号化」とかだと
区切りは重要ではないのでは

この場合，$\sum_x 2^{-l(x)} \le 1$ と $\sum_x 2^{-l(x)} \le N$

の違いそのものが小さい

$$\sum_x 2^{-l(x)-\log N} \le 1$$

問題その2： 符号化

確率が与えられたとして符号化を遂行できるか

m文字をブロック化 文字がK種類
abcaabcc aaabbaca bacccacc
m=8文字 (K=3)

$$K^m=\exp(m\log K)$$

実はさっきの「証明」の符号化法は半端の処理がベストでない
(ベストの方法 ⇒ハフマン符号化)
いずれにしても計算量がmの指数で発散

算術符号

独立＆確率が
(1/3,2/3)の場合

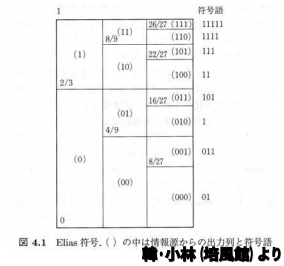


図 4.1 Elias 符号.（）の中は情報源からの出力列と符号語

岡・小林（培風館）より

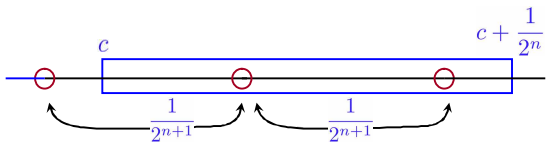
$P(x_i|x_{i-1})$

条件つき確率にしたがって
逐次的に
一個の列（ファイル）
に一個の
実数の区間
を対応させる

実数の区間が符号になる？

実数 ‥ 無限桁なので符号としては無意味

実数の区間 → 幅が広いほど
「簡単な2進小数」を含む



符号化

$0.0110001000000\cdots$

⇩

符号

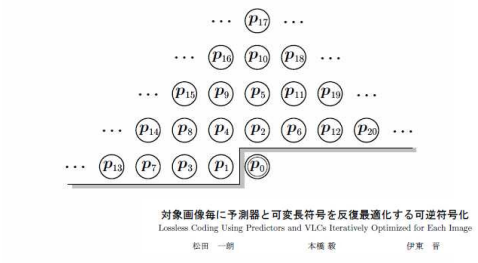
実際にやろうとすると超高精度の小数演算が必要
⇒ そこをなんとか処理して
効率のよい処理を
実現したのが算術符号

マルコフ連鎖を超えると？

条件付き確率の積で表示できる
ようなモデル（マルコフ連鎖，一般に巡回閉路を持たない有向グラフ上のモデル）
⇒ 算術符号にあっている

画像：条件つき確率

$P(p_0|p_1, p_2, \ldots, p_{17})$



対象画像毎に予測器と可変長符号を反復最適化する可逆符号化
Lossless Coding Using Predictors and VLCs Iteratively Optimized for Each Image
松田 一朗　本橋 毅　伊東 晋

問題その3　確率をどうやって知るか？

アルファベット26個の確率なら，たくさんの
文書から頻度を数えて．．でもよかった

大きな塊xを要素として確率P(x)を
考えるとなると，全く様相が変わってくる

統計科学との接点，　MDL原理，‥　後半へ！

Ⅳ エントロピーの意味

$$H = -\sum_x P(x) \log P(x)$$

情報理論　平均符号長

統計物理　エントロピー

カノニカル分布を前提として熱力学につながる
（↑の解釈は物理に限る）

カノニカル分布の場合

$$P(x) = \frac{\exp(-E(x)/T)}{Z}$$

$$\log P(x) = -E(x)/T + \log Z$$

カノニカル分布では温度に依存する定数

$$H = -\langle E \rangle / T + \log Z$$

$$F = -T \log Z$$

自由エネルギー

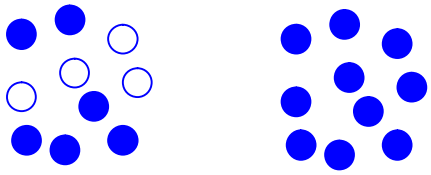
$$F = \langle E \rangle - TH$$

エントロピーSに一致

純粋に確率分布の性質として

$$H = -\sum_x P(x) \log P(x)$$

硬貨投げ



青の確率が0.51のときどっちが出やすい？

コインを区別するかどうかで違う

イジングでも同じ

javatest¥ising3.html

「確率の確率」という考え方

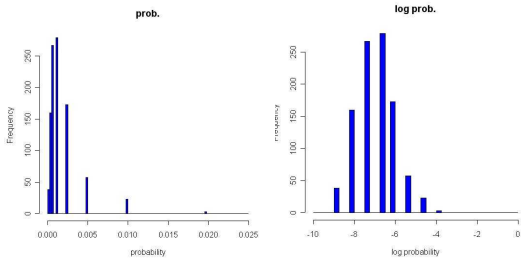
「ある確率で起こる」ことのどれかひとつが起きる確率分布

例　●が1/3 ○が2/3のとき
n回試行を行う　x=(●○○○●‥)

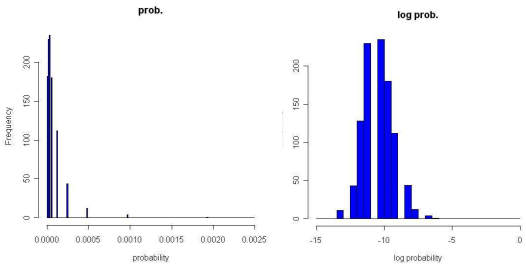
$$P(x) = \left(\frac{1}{3}\right)^m \left(\frac{2}{3}\right)^{n-m}$$

xの中の青丸の個数m

シミュレーション：n=8



シミュレーション：n=12



積の分布と和の分布

$$P(x) = \left(\frac{1}{3}\right)^m \left(\frac{2}{3}\right)^{n-m}$$ **対数正規分布**

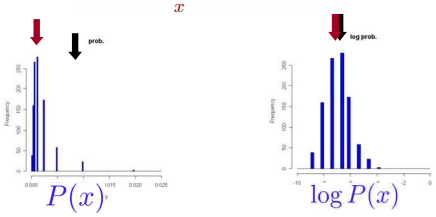
$$\log P(x) = n \log\left(\frac{2}{3}\right) + m \log\left(\frac{1}{2}\right)$$ **正規分布**

$$= m + (n-m) \log\left(\frac{2}{3}\right)$$

典型的な値

$P(x)$ ではなく $\log P(x)$ の相加平均

$$H = -\sum_x P(x) \log P(x)$$



エントロピーの意味

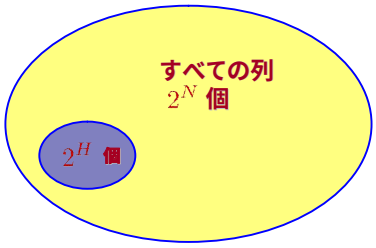
典型的な確率

$$2^{-H} = 2^{\sum_x P(x) \log P(x)}$$

頻繁に出る列の個数はおよそ

$$2^{H} = 2^{-\sum_x P(x) \log P(x)}$$

よくある絵



「確率」の基本

独立なら確率は **積**
関係なければ**和**になる量

確率論は積と和のなすドラマである

対数正規分布の例

@ 金融
利子が独立にランダムに変化
掛け算になる（複利だから）

@ 透明な板を重ねる

V. MDL原理 （最小記述長原理）

確率がわかってないときにどうするか

1 統計的に確率を推定
高次マルコフ 文脈木（可変長マルコフ）
PPM, CTW

2. ユニバーサル圧縮
Ziv-Lempel符号(LZ77,LZ78) gzip, lha
ブロックソート bzip2

統計的手法 対 情報圧縮固有の手法

統計的手法
確率を明示的な統計モデルで予測
⇒ 2つの分野の融合
「情報圧縮」の視野を拡大

固有の手法
広い意味では統計的予測と解釈できる
良い意味でのハッキング・スピリット
なかなか理屈だけでは勝てない（特に速度）

圧縮率

original	3407KB
lha level 4	1913KB
gzip	1592KB
bzip2	1480KB
lha level 7	1461KB
ppmz	1429KB
paq	1313KB

MDL原理

「確率が未知の場合の情報理論」
は統計学と情報理論の関係を再認識させ
「統計科学」の展開の一翼を担うこととなった

しかし，それだけではない
統計科学にとって根本的な問題が
情報圧縮の中にあらわれてくる **単純さ**

頻度を数える

確率がわかってないときにどうするか
⇒ とりあえず頻度を数えてみる

0100100101010000
0 1 0 0 1 0 0 1 0 1 0 1 0 0 0 0　1文字の頻度
01 00 10 01 01 01 00 00　2文字の頻度

もっとも単純な統計モデルともいえる

高次のマルコフを考えてもよいが本質的には同じ

相関と平均符号長

相関（非独立性）があれば
ブロック長大⇒理想符号長の平均は小さくなる

文字を2個まとめた場合について式で書くと

$$-\sum_{x_1,x_2,\cdots} P(x_1,x_2,\cdots)\log\frac{\prod_{i:odd}P(x_i,x_{i+1})}{\prod_{i:odd}P(x_i)P(x_{i+1})}$$

$$=-\sum_{x_1,x_2,\cdots}\left\{P(x_1,x_2,\cdots)\sum_{i:odd}\log\frac{P(x_i,x_{i+1})}{P(x_i)P(x_{i+1})}\right\}$$

$$=-\sum_{i:odd}\sum_{x_i,x_{i+1}}\left\{P(x_i,x_{i+1})\log\frac{P(x_i,x_{i+1})}{P(x_i)P(x_{i+1})}\right\}\le 0$$

どんどんブロックを大きくすると‥

0 1 0 0 1 0 0 1 0 1 0 1 0 0 0 0
01 00 10 01 01 01 00 00
010 010 010 101 000 0
0100 1001 0101 0000
01001 00101 01000 0
‥‥
0100100101010000

ん？

どんどん単調減少‥

最後に，ブロック長が圧縮するデータの長さに到達すると‥

P(X)　X：今あるデータなら　P(x)＝1
それ以外なら　P(x)＝0

$$-\sum_x P(x)\log P(x)=0$$

$$0\log 0+1\log 1=0$$

私的言語

「どんなデータ列の情報量もゼロ」

私がいいたいこと，たとえば
aajkkasssssajaa!!★!!
を1であらわすと「定義」
⇒ なんでも1ビット，いや0ビットで言える

※ 通じないけど
背後に想定した確率構造が共有されていない

辞書を忘れてはいけない

解読するためには辞書が必要

ブロック長 ＝ データ全体の長さ
辞書 ＝ もとのデータをそのまま含む

あきらかに無意味

MDL原理

辞書の長さ＋それで符号化した長さ

を最小にする

2段階符号化

それ以上の相関構造はとりこむべきでない

辞書＝確率モデル

「辞書」 ⇒ どのような確率（劣確率）を
もちいて符号化したかを表現

符号化の方式 （ハフマン，算術 ・・）
をいちばん最初に決めておけば

辞書＝確率モデル と考えてよい

辞書＝パラメータ

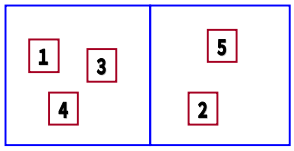
さらに確率モデルの族

$$P(x|\theta)$$

をはじめに決めてしまえば
辞書はパラメータ θ と同一視できる

ただし無限大の精度（実数）はいらない
⇒ 符号長が無限大になってしまう

2段階符号化の簡単な例



単純に考える

$$P(x) = \frac{1}{2^n}$$

$$l_e(x) = -n \log_e \frac{1}{2} = -\{\frac{n}{2} \log_e \frac{1}{2} + \frac{n}{2} \log_e \frac{1}{2}\}$$

2段階に分ける

$$P(m) = \frac{1}{n+1}$$

左の箱にm個

辞書
（パラメータ）
に相当

$$P(x|m) = \frac{1}{{}_nC_m} = \frac{m!(n-m)!}{n!}$$

そのm個がどれか

2段階符号化の符号長

$$-\log_e P(x|m) = -\log_e \frac{m!(n-m)!}{n!} \sim$$
$$-\left\{ m\log_e \frac{m}{n} + (n-m)\log_e \frac{n-m}{n} \right\} - \frac{1}{2}\log_e n$$

$$l_e(x) = -\log_e P(x|m) - \log_e P(m) \sim$$
$$-\left\{ m\log_e \frac{m}{n} + (n-m)\log_e \frac{n-m}{n} \right\} - \frac{1}{2}\log_e n + \log_e(n+1)$$

2段階符号化の得失

$$-\sum_x P(x)\log_e \frac{P(x)}{Q(x)} \le 0$$

$$l(x) \sim -n\left\{ \frac{m}{n}\log_e \frac{m}{n} + \frac{n-m}{n}\log_e \frac{n-m}{n} \right\} + \frac{1}{2}\log_e n$$

$$l(x) = -n\left\{ \frac{1}{2}\log_e \frac{1}{2} + \frac{1}{2}\log_e \frac{1}{2} \right\}$$ (単純考え) と比較する

ちょうど「左右半々」のときは単純考え
それ以外は2段階が漸近的に有利

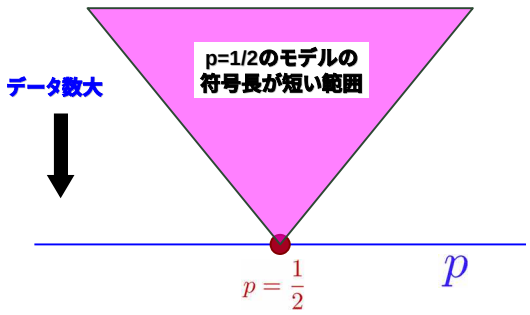
データ数が有限のとき

しかし，データ数が有限 (nが有限) であれば

$$\bar{l}(x) \sim -n\left\{ \frac{m}{n}\log_e \frac{m}{n} + \frac{n-m}{n}\log_e \frac{n-m}{n} \right\} + \frac{1}{2}\log_e n$$

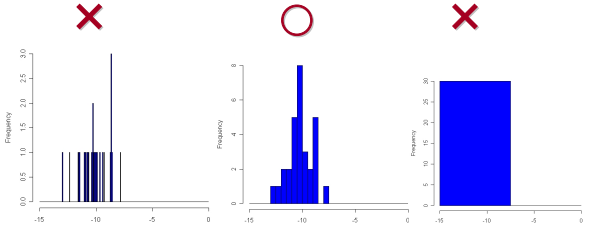
おまけの項の存在のために，データ数が少ない
場合には正確にm=n/2でなくても
P=1/2と決め打ちしたほうが有利になる

イメージ図



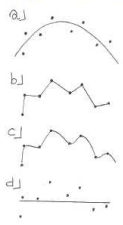
ヒストグラムの切り方

符号長という観点から考えることもできる



汎化 (generalization)

曲線をあてはめる



社会調査の解析

100項目 (YES, NO)

2^{100}通り

100次元の表の中身は0か1ばかりになりかねない

「どの人もこの世で唯一の人」

遺伝的アルゴリズムか何かでロボットに歩き方を学習させることを考える．うまく歩けなかった奴はボツにして，ましなやつを選抜してランダムに改良

この方法で「何が」学習されているのか

- 歩き方一般を学ぶ
- 訓練に使った廊下の構造を学ぶ
- その廊下のでこぼこやごみまで全部学ぶ

学習能力の大きいほど何でも丸暗記してしまい，ますます悪くなる可能性がある

「単純さ」の論理

なぜ単純なモデルが好まれる？
なぜ「規則」と「雑音」「偶然」に分ける？

MDL　情報圧縮の上でそれが有利だから
AIC　予測のためにそれが有利だから
仮説検定　主張する側に立証責任がある
ベイズ　事後確率が高い

赤池情報量規準 (AIC)　**本年度の京都賞**

■ 戦後の統計科学の転回点
(1970年代はじめ)

@ 簡単な数学と深い意味
cf. 木村資生の中立説
@ データ解析の現場からの貢献

■ 考え方の変革

@ 情報処理とは情報を捨てること
@ 規則と雑音の相対性
@ 統計学とはモデリングの学である

AICとMDL

・・・ は宿命のライバル，だったりするのだが
今日はそのあたりに深入りするのはやめて

(実際，これらから起きた流れは大きくひろがっていて単純な対決話はちょっともう古い)

「MDLはベイズ統計に近い」 という話を少し

MDLと事前分布

よく考えると「辞書」を圧縮するのにも
符号化を行ってよい

辞書が　$P(x|\theta)$
事前分布　$P(\theta)$

MDLの人たちもこのへんはいろいろ議論
さっきはうまくスルーできる例を選んだ

ベイズのモデル比較

$$P(M) \times \int P_M(y|x)P(x)dx$$

モデルの事後確率

さっきの場合（硬貨投げ）

$$\int P_M(y|x)P(x)dx$$

$$\int_0^1 p^m(1-p)^{n-m}dp = \frac{1}{n+1}\frac{1}{{}_nC_m}$$

ベータ関数の公式

2段階に分ける

$$P(m) = \frac{1}{n+1}$$

$$P(x|m) = \frac{1}{{}_nC_m} = \frac{m!(n-m)!}{n!}$$

$$P(x|m)P(m) = \frac{1}{n+1}\frac{1}{{}_nC_m}$$