九州大学 海洋システム工学専攻

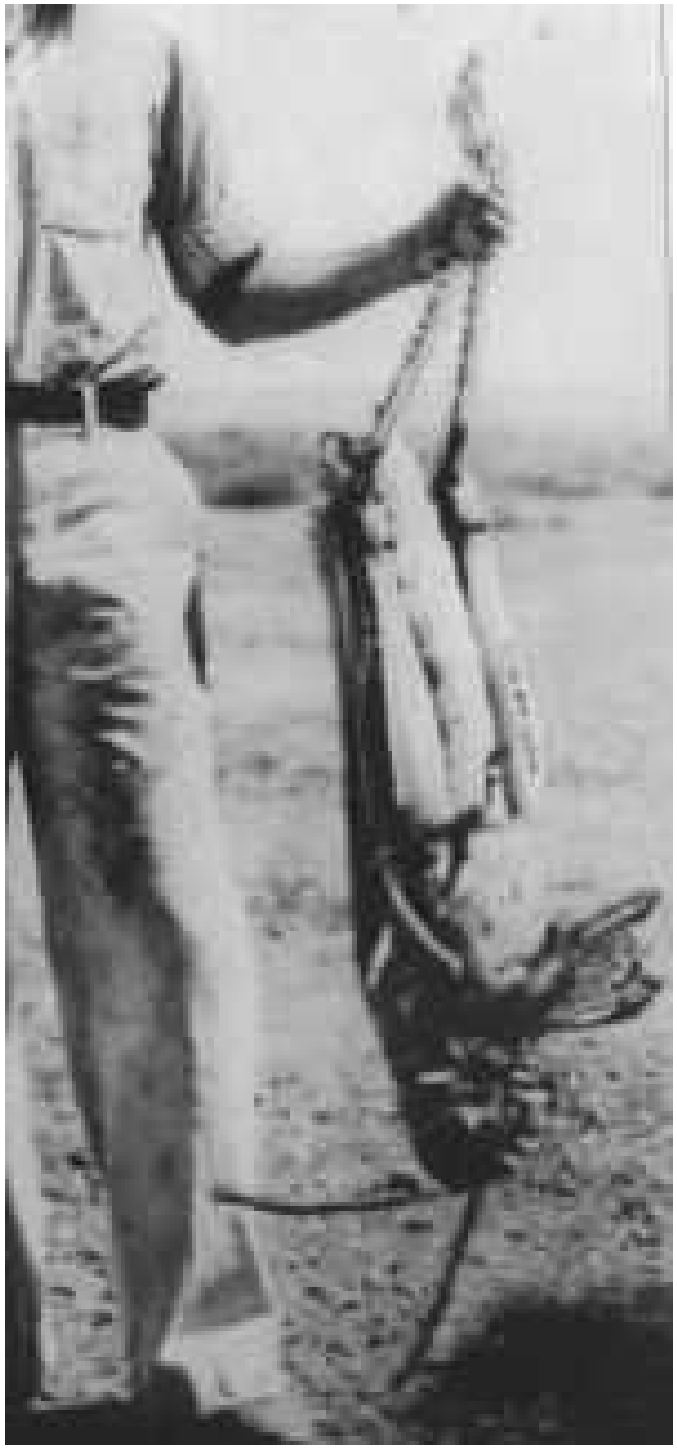
システム設計特論 （木村）

(3) 情報理論の基礎

「情報量」って何だろう？

授業の資料等は

http://sysplan.nams.kyushu-u.ac.jp/gen/index.html



情報理論の基礎

【情報とは何だろう？】

あることがらについての**知らせ**。
判断を下したり行動を起こしたりするために必要な、数々の媒体を介しての**知識**。
（広辞苑より）

【情報の定量化】

（1）情報が伝える事がらによって価値が異なる
　・必ず起こることが分かっている結果が知らされても、その情報に価値はない
　・めったに起きない現象が生じたことを知らされると、その情報には大きな価値がある
→ **情報の価値は、その情報が伝える事象の［　　］に依存し、［　　］が低いほど大きい**

（2）2つの独立な事象を伝えている情報は、それぞれを別々に伝える情報の価値の合計
→ 情報の価値は、伝えられる（独立な）事象に比例

（3）同じ情報でも、受け手によって価値が異なる
・一度読んでしまった新聞はただの紙だが、まだ読んでいない人にとっては価値がある
→ **受け手が知らない情報は、その受け手にとって価値が高い**

情報理論の基礎

【情報とは何だろう？】

あることがらについての**知らせ**。
判断を下したり行動を起こしたりするために必要な、数々の媒体を介しての**知識**。
（広辞苑より）

【情報の定量化】

（1）情報が伝える事がらによって価値が異なる
　・必ず起こることが分かっている結果が知らされても、その情報に価値はない
　・めったに起きない現象が生じたことを知らされると、その情報には大きな価値がある
→ **情報の価値は、その情報が伝える事象の 確率 に依存し、 確率 が低いほど大きい**

（2）2つの独立な事象を伝えている情報は、それぞれを別々に伝える情報の価値の合計
→ 情報の価値は、伝えられる（独立な）事象に比例

（3）同じ情報でも、受け手によって価値が異なる
・一度読んでしまった新聞はただの紙だが、まだ読んでいない人にとっては価値がある
→ **受け手が知らない情報は、その受け手にとって価値が高い**

起こり得る結果が有限であるような確率的現象を考える
ある結果が確率 P で生起する
この結果を知ったときに得る情報の価値を I(P) で表す

（1）確率の小さい結果が生じたとき、それを伝える情報の価値は大きい
→ I(P) は P の単調減少関数

（2）2つの互いに独立な事象AとBの生起確率はそれぞれ Pa および Pb
AとBをまとめて一つの確率事象とするなら、確率は Pa × Pb
→ I(Pa × Pb) = I(Pa) + I(Pb)

以上の条件を満たす連続関数 I(P) は？

確率 0.5 の結果を知ったときに得る情報の価値

起こり得る結果が有限であるような確率的現象を考える
ある結果が確率 P で生起する
この結果を知ったときに得る情報の価値を I(P) で表す

（1）確率の小さい結果が生じたとき、それを伝える情報の価値は大きい
→ I(P) は P の単調減少関数

（2）2つの互いに独立な事象AとBの生起確率はそれぞれ Pa および Pb
AとBをまとめて一つの確率事象とするなら、確率は Pa × Pb
→ I(Pa × Pb) = I(Pa) + I(Pb)

以上の条件を満たす連続関数 I(P) は？

$$I(P) = -\log_2 P$$

確率 0.5 の結果を知ったときに得る情報の価値

$$I(0.5) = -\log_2 0.5 = 1 \quad (\text{bit})$$

復習： 対数関数の性質

$$\log xy = \log x + \log y$$

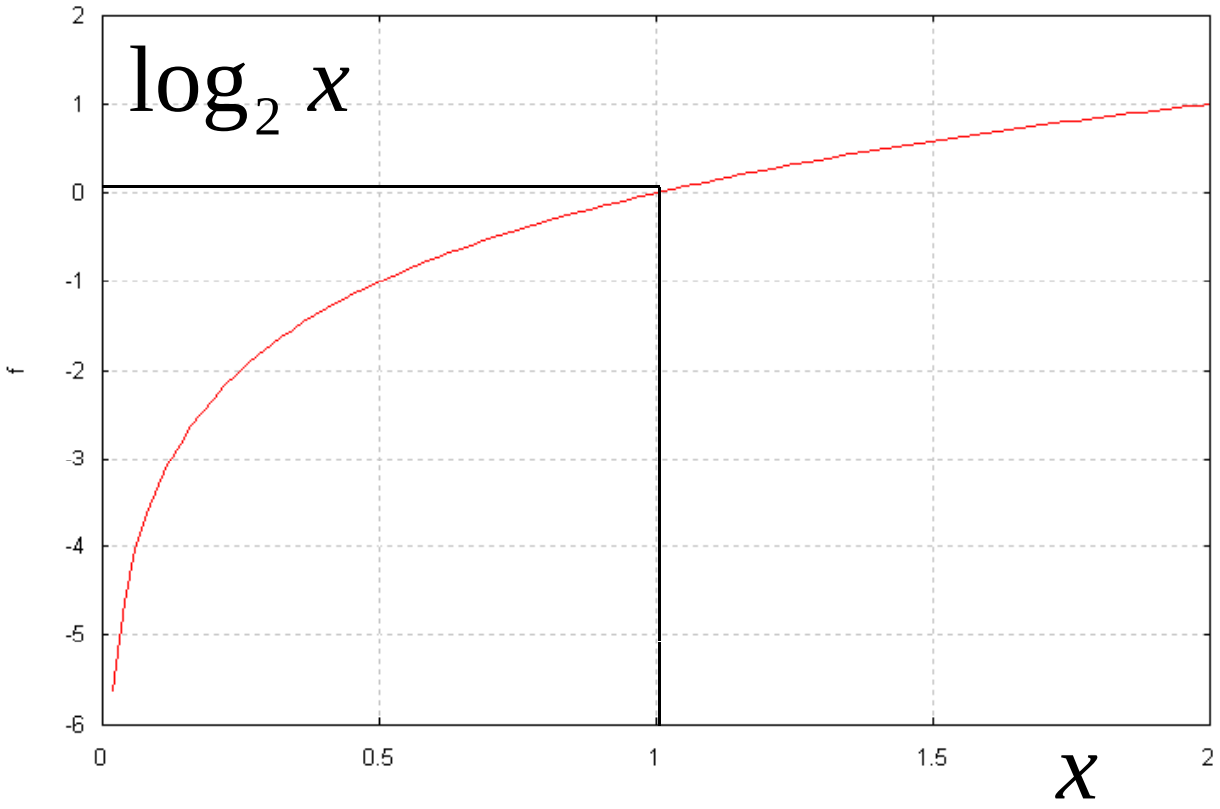
$$\log\left(\frac{x}{y}\right) = \log x - \log y$$

$$\log x^y = y \log x$$

$$\log 1 = 0$$

$$\log_2 2 = 1$$

$$\log_2 x = \frac{\log_e x}{\log_e 2}$$



情報量（平均情報量）: Information

A国では「晴」「雨」の情報は
どちらも1bitの価値

例） 天気についての情報
A国では　晴れの割合　1／2、　雨の割合　1／2
B国では　晴れの割合　99／100、雨の割合　1／100

B国では「晴」の情報の
価値は 0.0145 bit しかない

B国では「雨」の情報の
価値は 6.64 bit もある

両国の天気情報の価値の期待値を計算すると

A国　$\bar{I}_A = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1 \quad (\text{bit})$

B国　$\bar{I}_B = -\frac{99}{100}\log_2\frac{99}{100} - \frac{1}{100}\log_2\frac{1}{100} = 0.0144 + 0.0664 = 0.0808 \quad (\text{bit})$

B国での天気のニュースの情報量は、A国の8%程度

情報量（平均情報量）
単位：

情報量（平均情報量）: Information

A国では「晴」「雨」の情報は
どちらも1bitの価値

例） 天気についての情報
A国では　晴れの割合　1／2、　雨の割合　1／2
B国では　晴れの割合　99／100、雨の割合　1／100

B国では「晴」の情報の
価値は 0.0145 bit しかない

B国では「雨」の情報の
価値は 6.64 bit もある

両国の天気情報の価値の期待値を計算すると

A国　$\bar{I}_A = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1 \quad (\text{bit})$

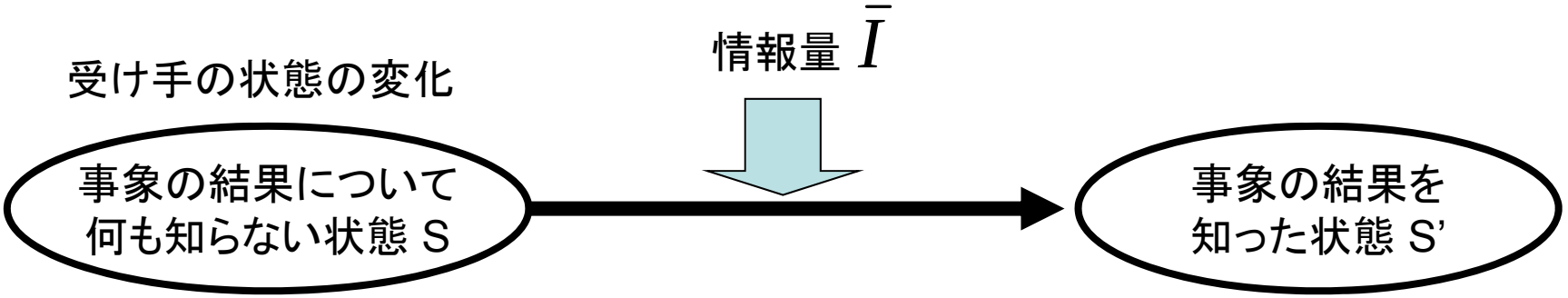
B国　$\bar{I}_B = -\frac{99}{100}\log_2\frac{99}{100} - \frac{1}{100}\log_2\frac{1}{100} = 0.0144 + 0.0664 = 0.0808 \quad (\text{bit})$

B国での天気のニュースの情報量は、A国の8%程度

$$\bar{I} = -\sum_{i=1}^{M} P_i \log_2 P_i$$

情報量（平均情報量）
単位：bit

情報量とエントロピー：あいまいさを表す尺度



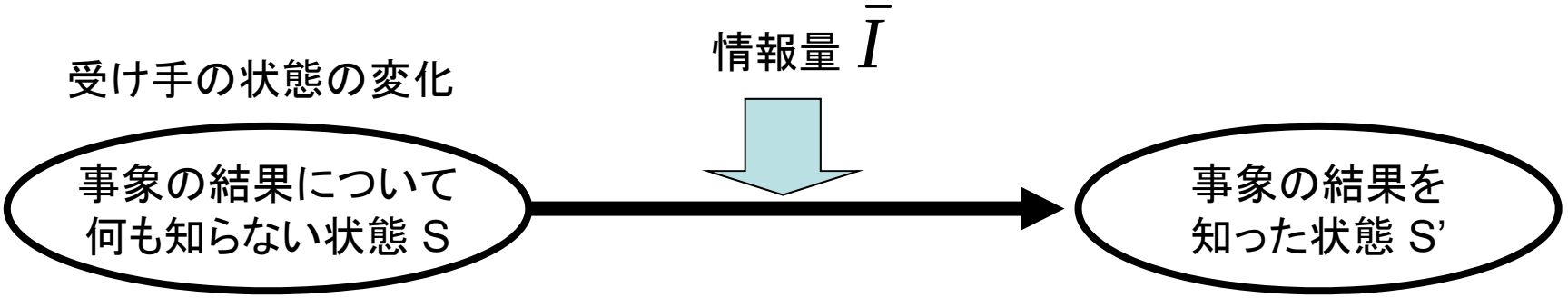
「情報量」＝その情報を受け取ることによるエントロピーの減少

$$H(S) - \bar{I} = H(S')$$

エントロピー　　（単位：bit）

エントロピーは、全ての事象が等確率 1/Mで発生するとき最大値　$\log_2 M$

情報量とエントロピー：あいまいさを表す尺度



「情報量」＝その情報を受け取ることによるエントロピーの減少

$$H(S) - \bar{I} = H(S')$$

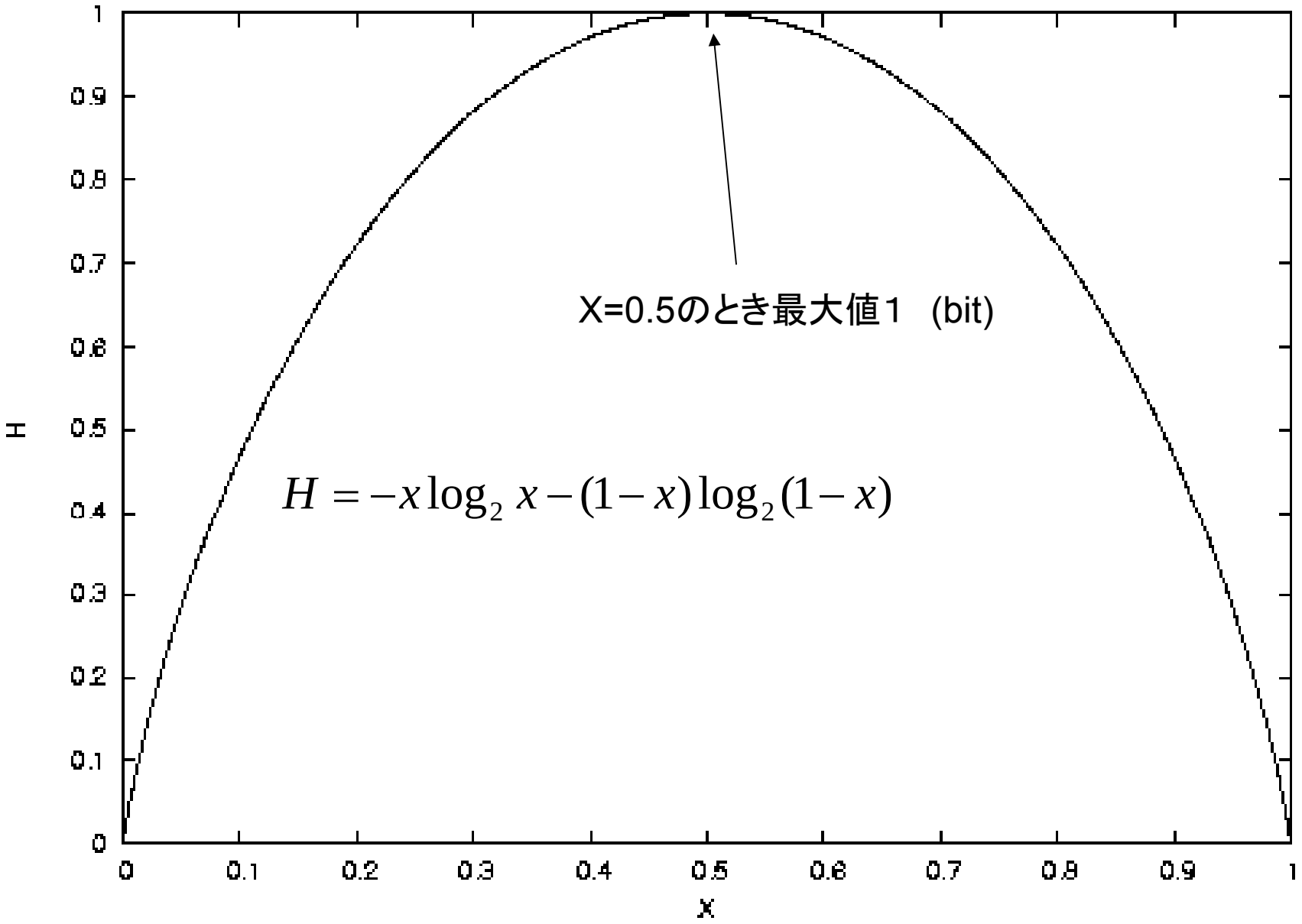
エントロピー

$$H(S) = -\sum_{i=1}^{M} P_i \log_2 P_i$$

（単位：bit）

エントロピーは、全ての事象が等確率 1/Mで発生するとき最大値　$\log_2 M$

確率 x で表、確率 1-x で裏が出る事象のエントロピー



情報量（エントロピー）と伝送データ圧縮

データ伝送と符号化

例） アルファベットで記述された文書を、0と1の信号を組み合わせて伝送する
伝送コストは、信号のビット数に比例する

アルファベットにどのような符号を割り当てるべきか？

A＝00000
B＝00001
C＝00010
D＝00011
…

固定長の符号
（符号長5bit）

1文字あたりの符号長は常に同じ
26文字を5 bit で表現すると
少し冗長で無駄

A＝1　　符号長1（bit）
B＝01　　符号長2（bit）
C＝001　　符号長3（bit）
D＝0001　　符号長4（bit）
…

可変長の符号

符号長の短いものを
頻度の高い文字へ割り当てれば、
1文字あたりの平均符号長を
短くできる

英文における文字の出現確率

文字	確率		文字	確率		文字	確率
A	8.29%		J	0.21%		S	6.33%
B	1.43%		K	0.48%		T	9.27%
C	3.68%		L	3.68%		U	2.53%
D	4.29%		M	3.23%		V	1.03%
E	12.08%		N	7.16%		W	1.62%
F	2.20%		O	7.28%		X	0.20%
G	1.71%		P	2.93%		Y	1.57%
H	4.54%		Q	0.11%		Z	0.09%
I	7.16%		R	6.90%			

各文字が等確率で現れる場合、1文字あたりのエントロピーは

$$\log_2 26 = 4.70 \quad \text{(bit)}$$

各文字が表にしたがって現れる場合、1文字あたりのエントロピーは

$$H(S) = -\sum_{i=1}^{26} P_i \log_2 P_i = 4.17 \quad \text{(bit)}$$

これ以上は
データ圧縮
できない！

平均符号長の下限値は [　　] で与えられる （情報源符号化定理）

英文における文字の出現確率

文字	確率		文字	確率		文字	確率
A	8.29%		J	0.21%		S	6.33%
B	1.43%		K	0.48%		T	9.27%
C	3.68%		L	3.68%		U	2.53%
D	4.29%		M	3.23%		V	1.03%
E	12.08%		N	7.16%		W	1.62%
F	2.20%		O	7.28%		X	0.20%
G	1.71%		P	2.93%		Y	1.57%
H	4.54%		Q	0.11%		Z	0.09%
I	7.16%		R	6.90%			

各文字が等確率で現れる場合、1文字あたりのエントロピーは

$$\log_2 26 = 4.70 \quad \text{(bit)}$$

各文字が表にしたがって現れる場合、1文字あたりのエントロピーは

$$H(S) = -\sum_{i=1}^{26} P_i \log_2 P_i = 4.17 \quad \text{(bit)}$$

これ以上は
データ圧縮
できない！

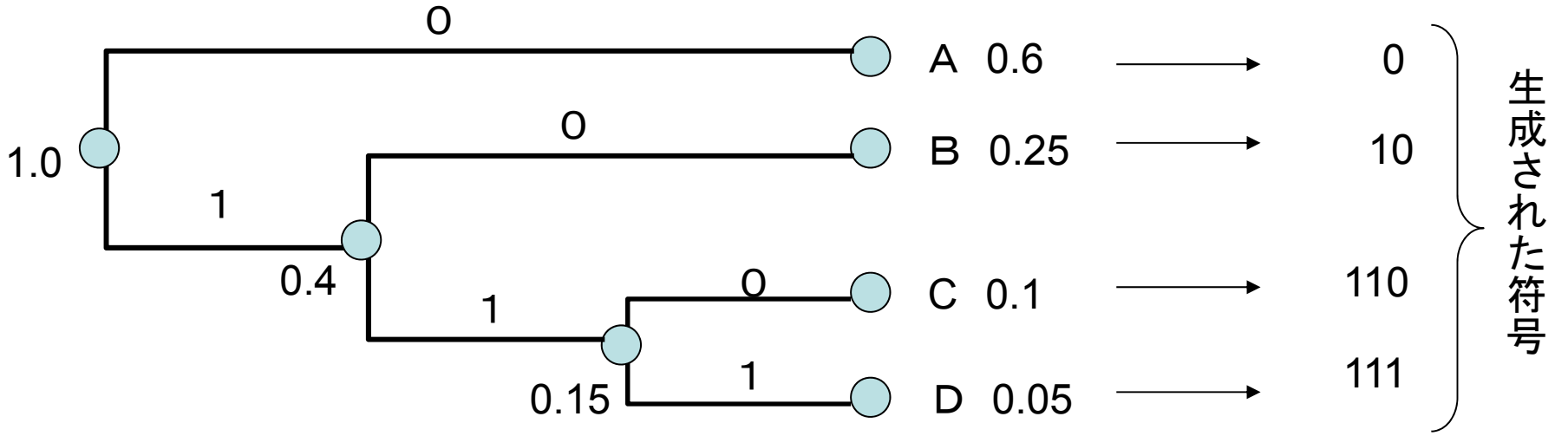
平均符号長の下限値は エントロピー で与えられる （情報源符号化定理）

平均符号長を最小化する符号化方法：ハフマン符号

2元ハフマン符号構成法

（1） 各アルファベットに対応する葉ノードを作り、その文字の発生確率を付加
（2） 発生確率の最も小さい2つの葉ノードに対し、1つの節点ノードを作りそれぞれをつなぐ。
この2本の枝の一方に「0」、他方には「1」を割り当てる。
新たに作ったノードには、つなげた2つのノードの確率の和を付加し、
これを新しい葉ノードとする。
（3） 葉ノードが1枚しか残っていなければ処理を終了、さもなければ（2）へ戻る

【例】A，B，C，Dがそれぞれ確率 0.6、0.25、0.1、0.05　で発生するとき

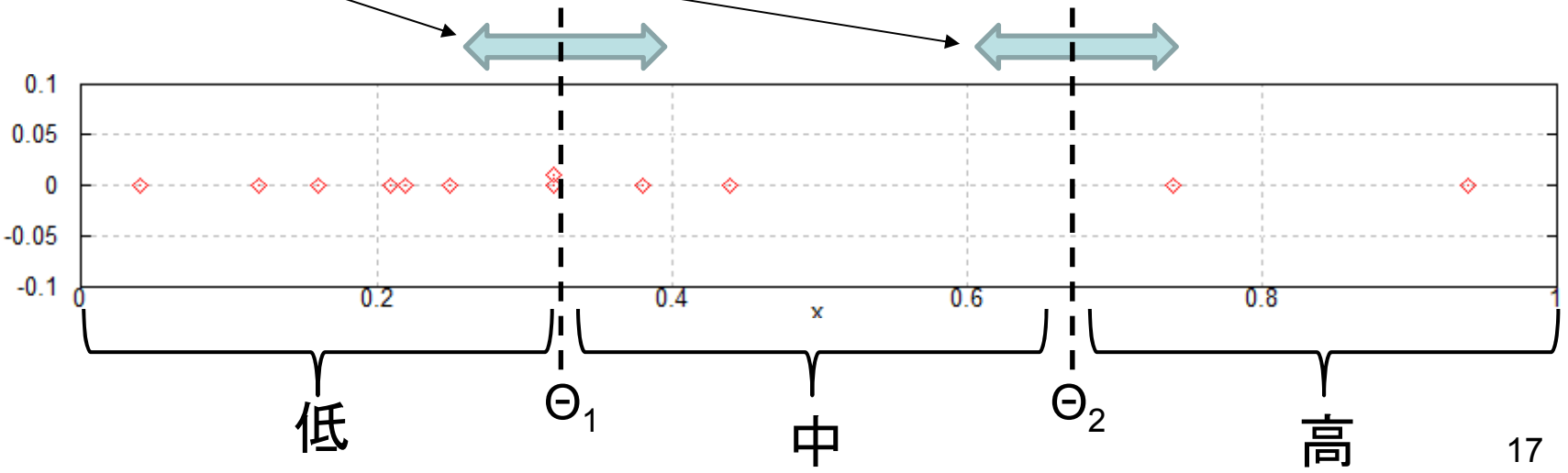


エントロピー 1.49 (bit)　＜　平均符号長は 1.55 (bit)

データの量子化（クラスタリング・教師無し学習）

試行	1	2	3	4	5	6	7	8	9	10	11	12
Xの観測値	0.21	0.74	0.32	0.25	0.12	0.38	0.32	0.22	0.04	0.44	0.94	0.16

【例】
表のように観測される確率事象において，確率変数 x の観測値は 0 ≦ x ＜ 1 の連続値だが，あるしきい値 Θ_1, Θ_2を設定して、0 ≦ x ＜ Θ_1 のとき「低」，
Θ_1 ≦ x ＜ Θ_2 のとき「中」,Θ_2 ≦ x < 1 のとき「高」, という表示を行うことを考える.
このとき，「低」「中」「高」による表示の**平均情報量を最大化**するには，
しきい値 Θ_1, Θ_2 をどのように設定したら良いか？



データの量子化（クラスタリング・教師無し学習）

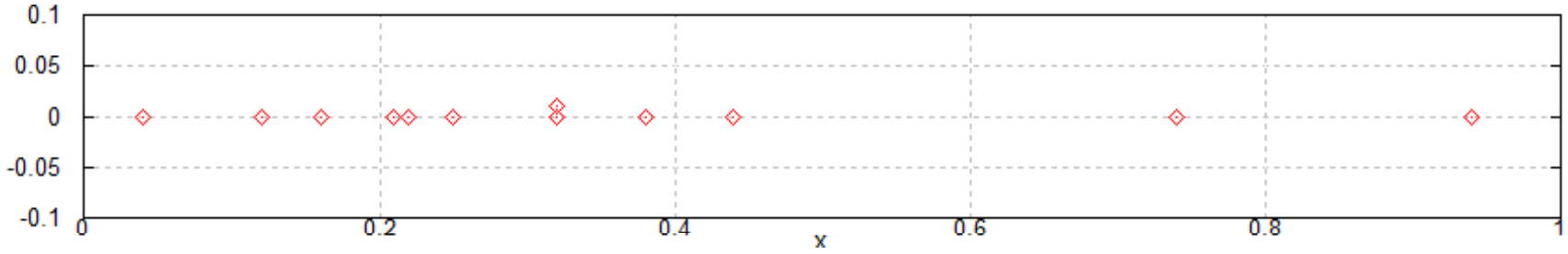
「低」に分類される事象の確率を　P_1
「中」に分類される事象の確率を　P_2
「高」に分類される事象の確率を　P_3　と表すと、全表示の平均情報量 I は

$$I = - P_1 \log_2 P_1 - P_2 \log_2 P_2 - P_3 \log_2 P_3$$

ただし　$P_1 + P_2 + P_3 = 1$ である。
よって等式制約条件下での極値問題になるので、
ラグランジュの未定乗数法によって I が最大になる P_1, P_2, P_3 を求めると

空間がn次元の場合は？
次回「クラスタリング」で説明

データは12個なので、これらを



データの量子化（クラスタリング・教師無し学習）

「低」に分類される事象の確率を　P_1
「中」に分類される事象の確率を　P_2
「高」に分類される事象の確率を　P_3　と表すと、全表示の平均情報量 I は

$$I = - P_1 \log_2 P_1 - P_2 \log_2 P_2 - P_3 \log_2 P_3$$

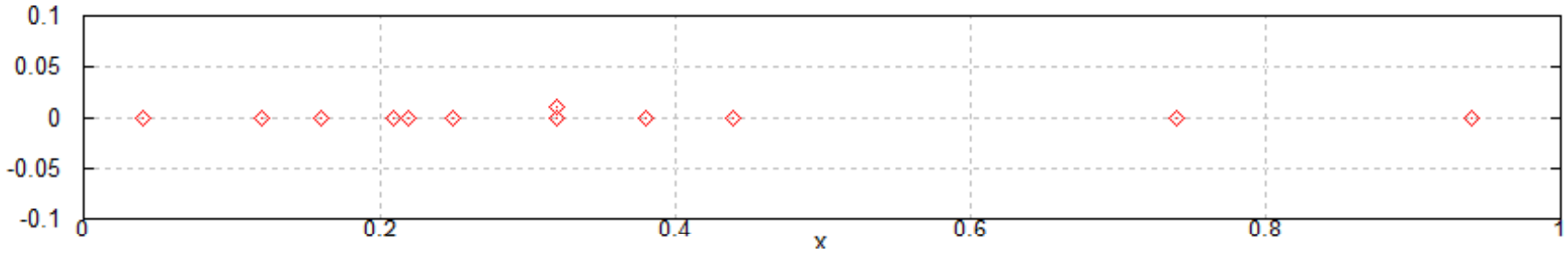
ただし　$P_1 + P_2 + P_3 = 1$ である。
よって等式制約条件下での極値問題になるので、
ラグランジュの未定乗数法によって I が最大になる P_1, P_2, P_3 を求めると

$P_1 = P_2 = P_3 = 1／3$　である。

空間がn次元の場合は？
次回「クラスタリング」で説明

データは12個なので、これらを

ちょうど4個ずつ3等分するようにしきい値Θ_1, Θ_2 を設定すれば良い。



データの量子化（クラスタリング・教師無し学習）

「低」に分類される事象の確率を　P_1
「中」に分類される事象の確率を　P_2
「高」に分類される事象の確率を　P_3　と表すと、全表示の平均情報量 I は

$$I = - P_1 \log_2 P_1 - P_2 \log_2 P_2 - P_3 \log_2 P_3$$

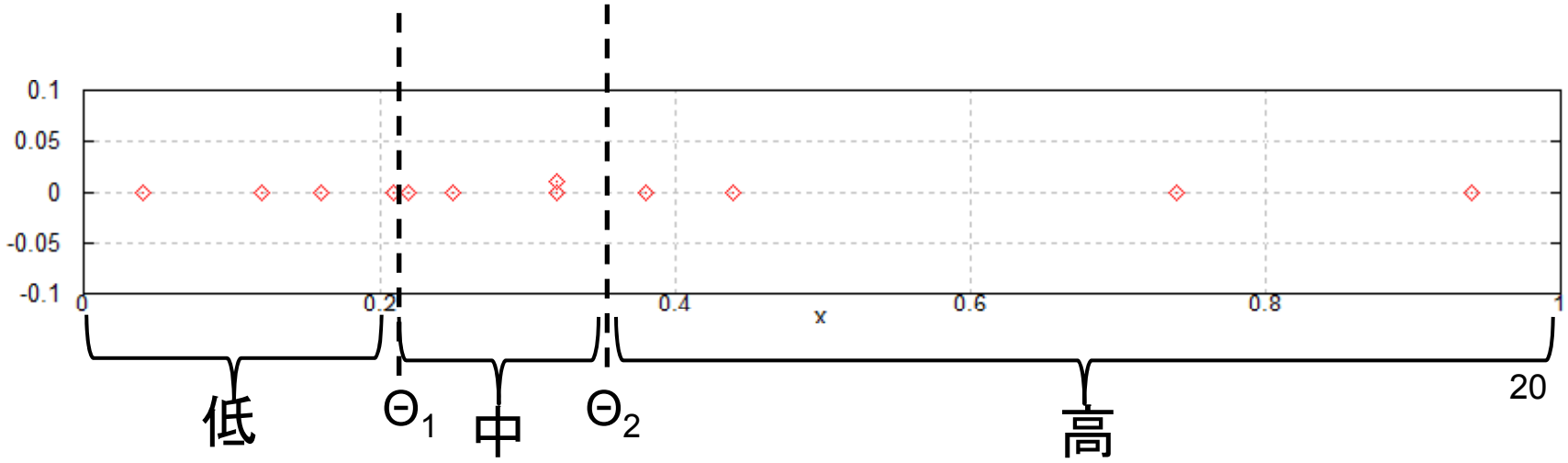
ただし　$P_1 + P_2 + P_3 = 1$ である。
よって等式制約条件下での極値問題になるので、
ラグランジュの未定乗数法によって I が最大になる P_1, P_2, P_3 を求めると

$P_1 = P_2 = P_3 = 1／3$　である。

空間がn次元の場合は？
次回「クラスタリング」で説明

データは12個なので、これらを

ちょうど4個ずつ3等分するようにしきい値Θ_1, Θ_2 を設定すれば良い。



姓名判断：　人の姓名からその人の性格や人生の趨勢、適職、恋愛の傾向、結婚運・家庭運、かかりやすい病気など、一般に運勢として総称される事柄について解釈を与える占いの手法（Wikipediaより）

人の姓名で使用する文字の画数から5つの格数を算出し、
それらに与えられた伝統的・経験的な解釈に基づいて解釈を行うものが主流

大量のサンプルから　[姓名]　→　[運勢]　を説明する単純なルールを形成

偉人・著名人・長寿老人・犯罪者・病気などで不幸に見舞われた人の名前を収集・**分類**

家相・風水：土地や家の間取りなどの相（見た目、ありさま）、またはそれによって住人の運勢をみる占術（Wikipediaより）

井戸・台所・風呂場・便所・廊下・階段などの設置位置関係や方位の吉凶を判断

大量のサンプルから　[家屋の間取]　→　[吉凶]　を説明する単純なルールを形成

偉人・著名人・長寿老人・犯罪者・病気などで不幸に見舞われた人等を輩出した家屋や火事等が発生した家屋の情報を収集・**分類**

決定木(分類木)の学習への応用　判別問題

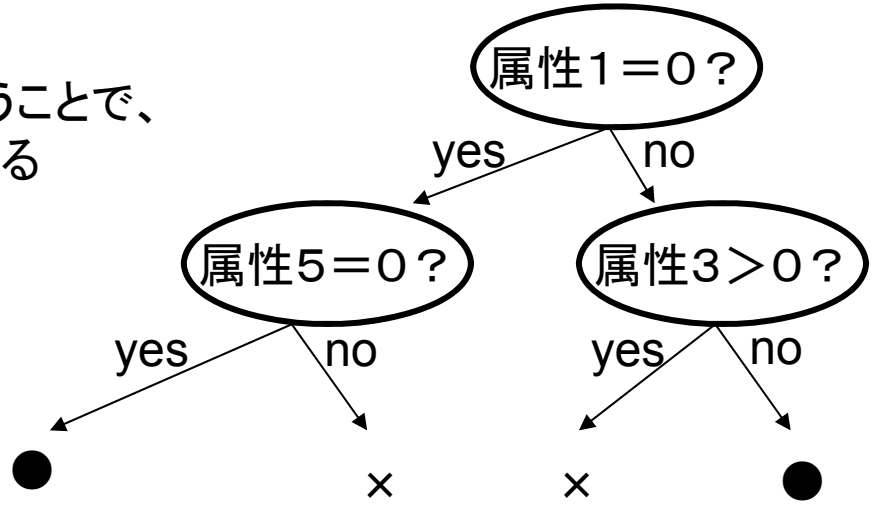
膨大な事例データから、結果を説明する**単純なルール**を見つける　→　データ解析

事例	属性1	属性2	属性3	属性4	属性5	結果
1	0	0	0.4	0.3	0	●
2	0	0	0.6	0.2	1	×
3	1	0	0.2	0.3	1	●
4	0	1	0.3	0.9	1	×
5	1	0	0.55	0.8	0	●
6	1	0	0.1	0.4	1	●
7	0	1	0.2	0.7	0	×
8	0	1	0.9	0.2	1	×
9	0	0	0.8	0.6	1	●
10	1	1	0.3	0.8	0	×

決定木(分類木)とは？

属性x の性質についての逐次的なテストを行うことで、
最終的に事例が属するクラス(結果)を推測する

テスト　　＝ ノード(node)
テスト結果　＝　枝の分岐



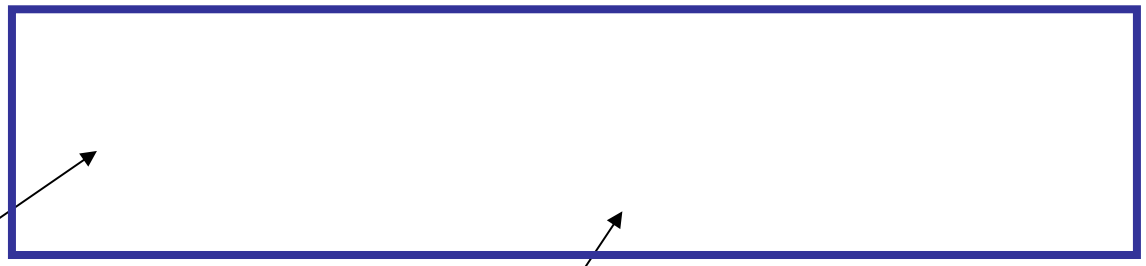
どのようにして単純な木を生成するか？

一般的な決定木の生成方法

全てのデータが同じクラス
or
全てのテストが同じ結果

（1）木の根となるノードに、全学習データDを対応付ける。
（2）・ノードに対応付けられたデータが、停止条件を満たす場合
そのノードに対応付けられたデータが属するクラスについて多数決をとり、
そのノードのクラスを決める（そのノードは決定木の葉ノードとなる）
・さもなければ、ノードに対応付けられたデータに対して適用可能なテストを選択し、
その結果によって子ノードを生成する。そのテスト結果に応じて子ノードへデータ
を分割して対応付ける。
（3）全てのノードで木の生成が停止するまで（2）より繰返す。

C4.5 （Quinlan 1986）の方法：
テストによる分割前と分割後で、データ内のクラスのエントロピーの減少が最大になるテスト
を選択
テータを2分割する場合、分割前のデータをD0、分割後のデータをD1, D2、
データの個数を |Di| 、データDiのエントロピーを H(Di) と表すと、エントロピーの減少量は



この式が最大になる
テストを選んでいく

分割前のエントロピー

分割後のエントロピーの加重平均

一般的な決定木の生成方法

全てのデータが同じクラス
or
全てのテストが同じ結果

（1）木の根となるノードに、全学習データDを対応付ける。
（2）・ノードに対応付けられたデータが、停止条件を満たす場合
そのノードに対応付けられたデータが属するクラスについて多数決をとり、
そのノードのクラスを決める（そのノードは決定木の葉ノードとなる）
・さもなければ、ノードに対応付けられたデータに対して適用可能なテストを選択し、
その結果によって子ノードを生成する。そのテスト結果に応じて子ノードへデータ
を分割して対応付ける。
（3）全てのノードで木の生成が停止するまで（2）より繰返す。

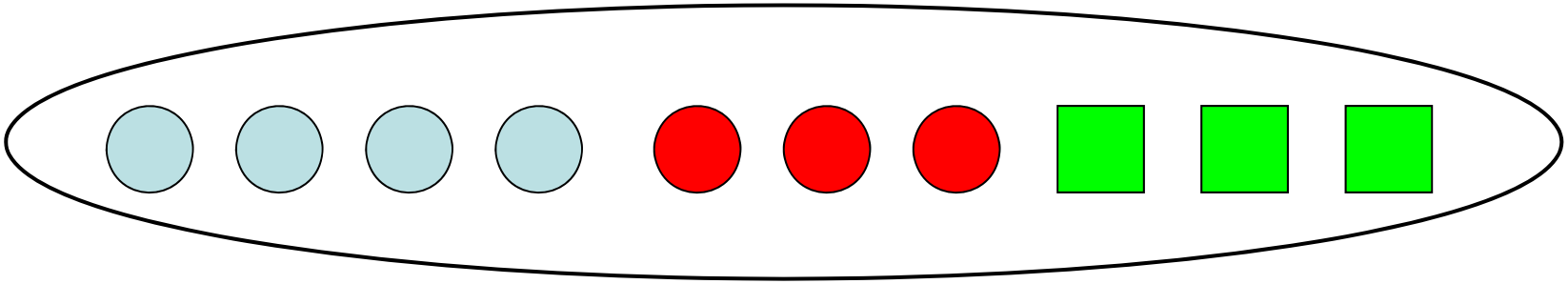
C4.5 （Quinlan 1986）の方法：
テストによる分割前と分割後で、データ内のクラスのエントロピーの減少が最大になるテストを選択
テータを2分割する場合、分割前のデータをD0、分割後のデータをD1, D2、
データの個数を |Di| 、データDiのエントロピーを H(Di) と表すと、エントロピーの減少量は

$$H(D_0) - \left(\frac{|D_1|}{|D_0|} H(D_1) + \frac{|D_2|}{|D_0|} H(D_2) \right)$$

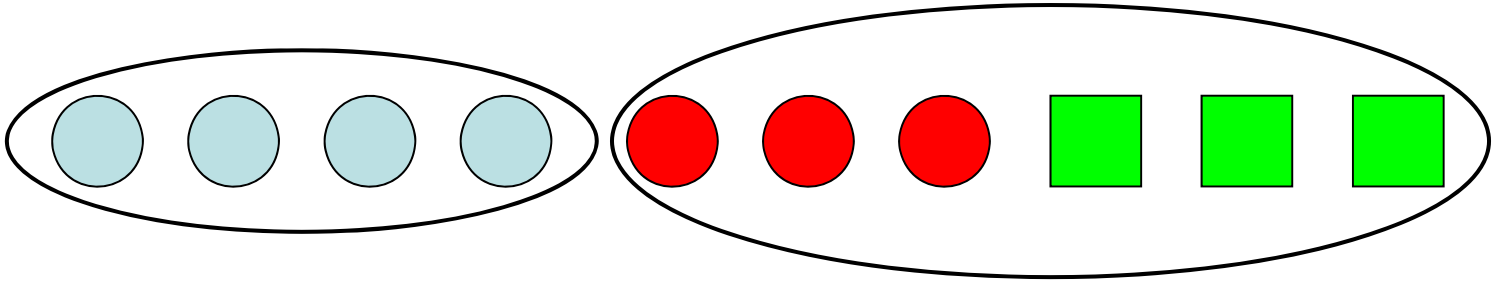
この式が最大になるテストを選んでいく

分割前のエントロピー

分割後のエントロピーの加重平均



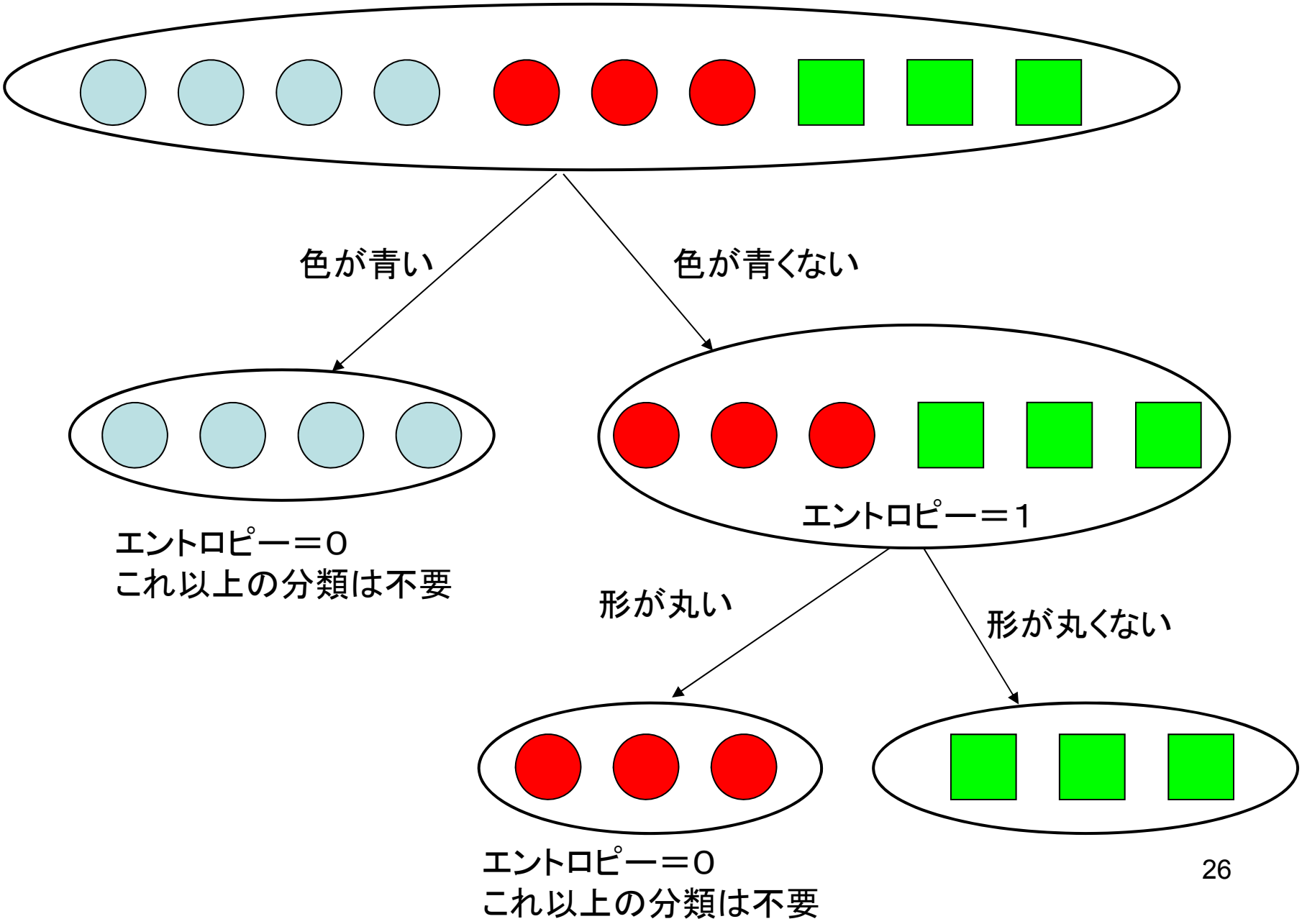
分割前のエントロピー $$\left(-\frac{4}{10}\log_2\frac{4}{10}\right)+\left(-\frac{3}{10}\log_2\frac{3}{10}\right)+\left(-\frac{3}{10}\log_2\frac{3}{10}\right)=1.57 \quad \text{(bit)}$$



2分割後のエントロピー $$\frac{4}{10}\left(-\frac{4}{4}\log_2\frac{4}{4}\right)+\frac{6}{10}\left(\left(-\frac{3}{6}\log_2\frac{3}{6}\right)+\left(-\frac{3}{6}\log_2\frac{3}{6}\right)\right)=0.6 \quad \text{(bit)}$$

3分割・4分割なども同様に計算できる

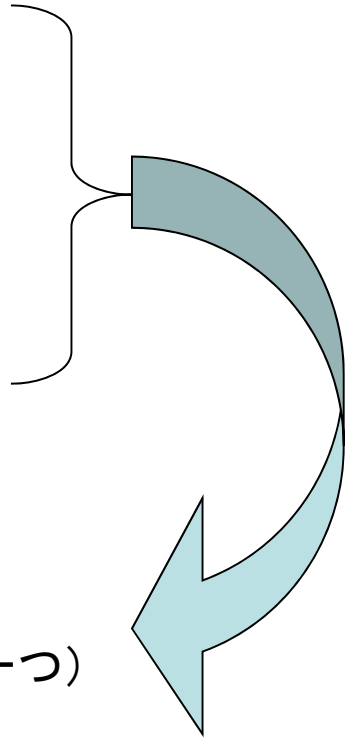
決定木（分類木）の生成例


色が青い
色が青くない
エントロピー＝０
これ以上の分類は不要
エントロピー＝１
形が丸い
形が丸くない
エントロピー＝０
これ以上の分類は不要

決定木の評価： 好ましい「知識表現」とは？

●誤り率の少ないものが好ましい
→ 葉ノードに全データが対応： 過学習（over-fitting）
未知のデータに対する推定が難しくなる

●小さくて簡単な構造の木が好ましい（記述長）



両方の評価のバランスをとる評価規範：
MDL（Minimum Description Length）
（データを説明するための一般的な確率モデルの評価方法の一つ）

「モデルの記述長」＋「モデルに対するデータの記述長」 ⇒ 最小化

（木の大きさに相当）

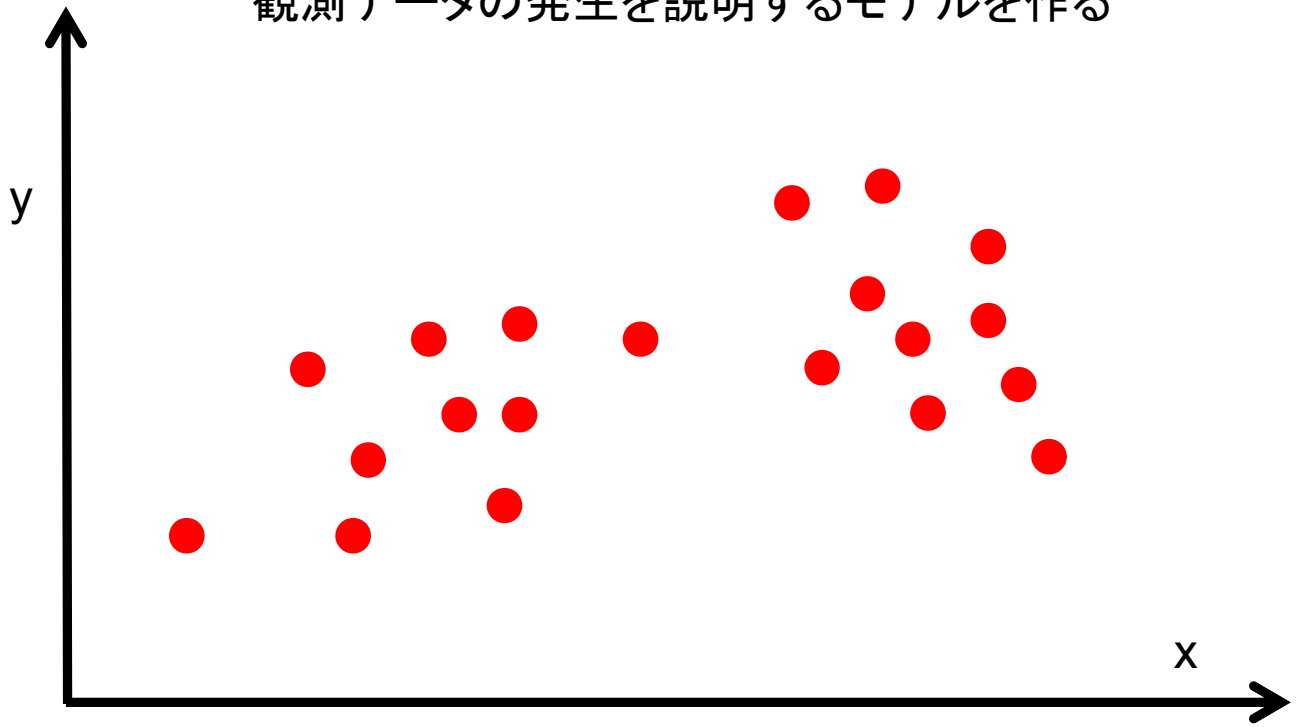
推定パラメータに対するデータの記述長
（各葉ノードにおけるデータのエントロピーに相当）
すなわち正答率のようなもの

この他、類似の評価規範としてAICなどがある： 詳細は専門書を参照のこと

最尤推定と確率モデル構築：　究極の「データ解析」

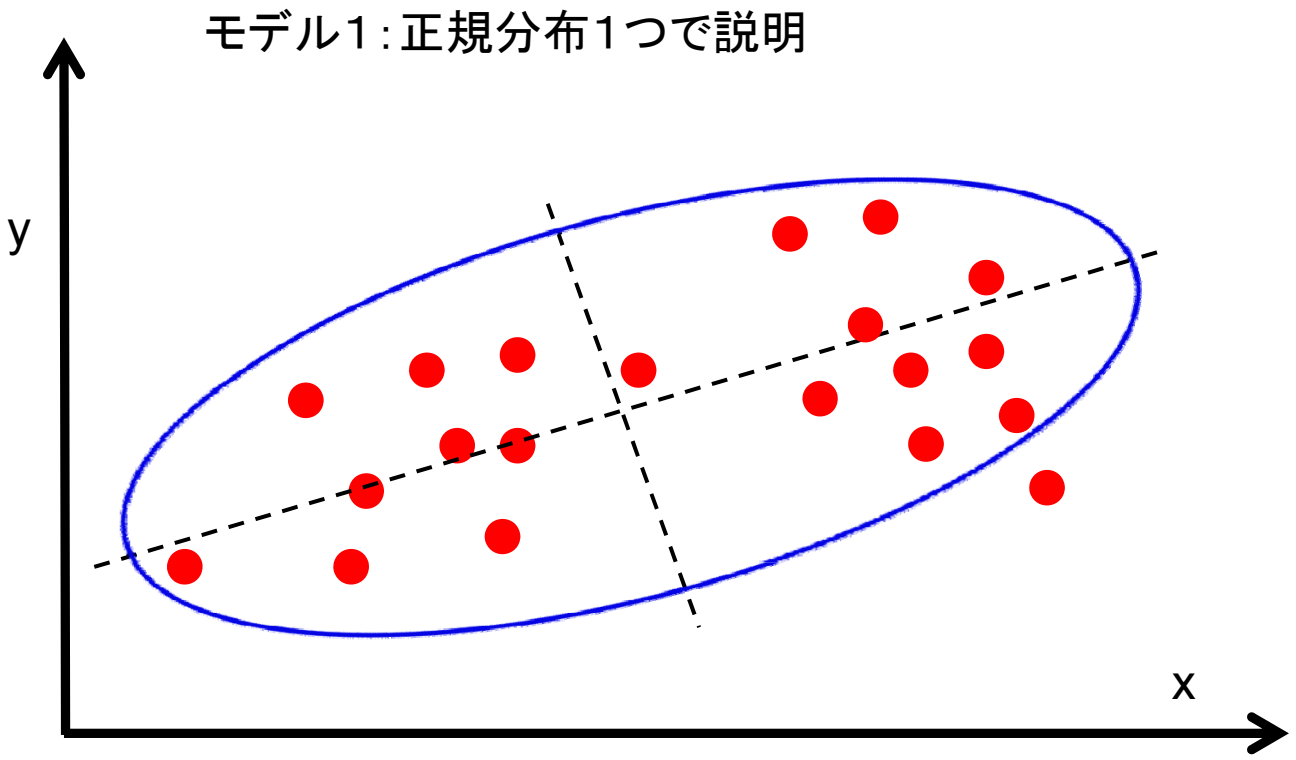
膨大な事例データから、結果を説明する**単純なルール**を見つける

観測データの発生を説明するモデルを作る



最尤推定と確率モデル構築：　究極の「データ解析」

膨大な事例データから、結果を説明する**単純なルール**を見つける



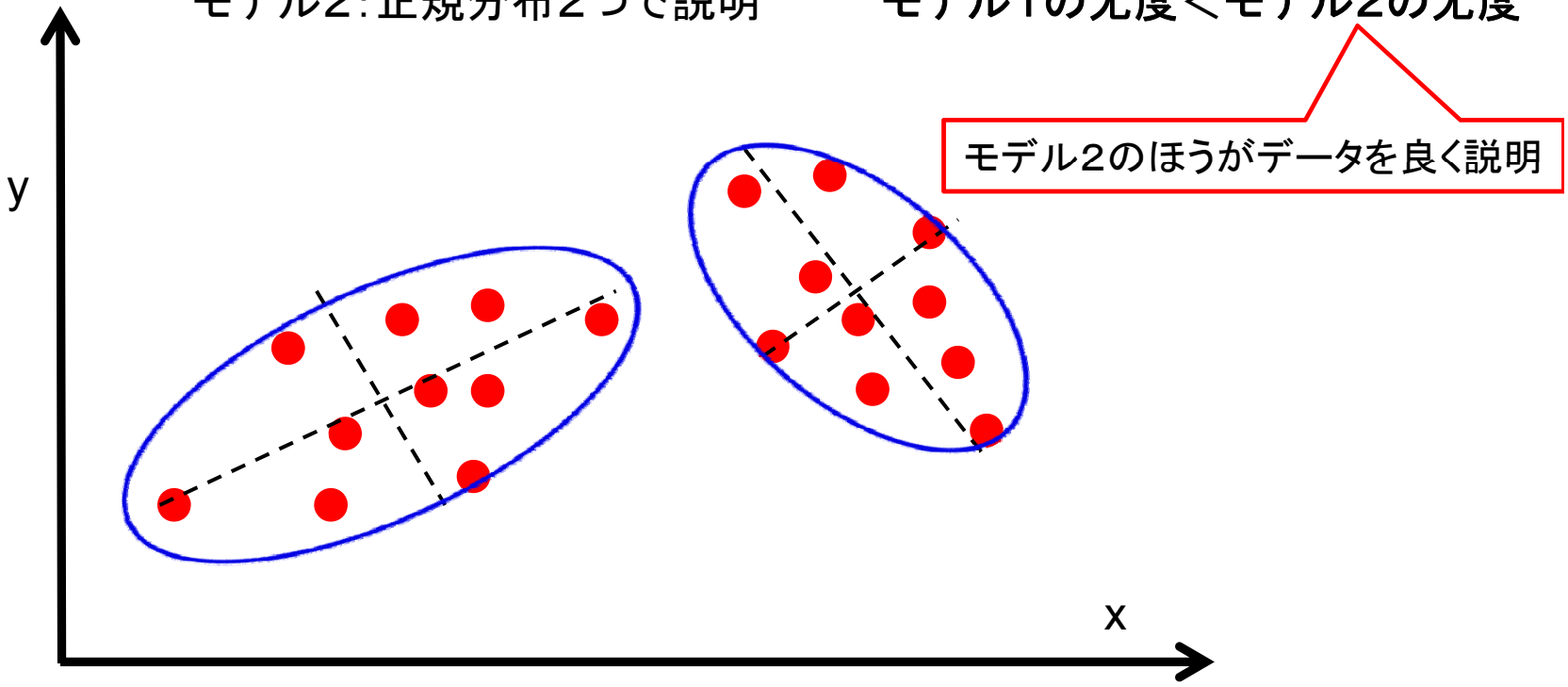
最尤推定と確率モデル構築：　究極の「データ解析」

膨大な事例データから、結果を説明する**単純なルール**を見つける

モデルから観測データが発生する確率

モデル2：正規分布2つで説明　　**モデル1の尤度＜モデル2の尤度**

モデル2のほうがデータを良く説明



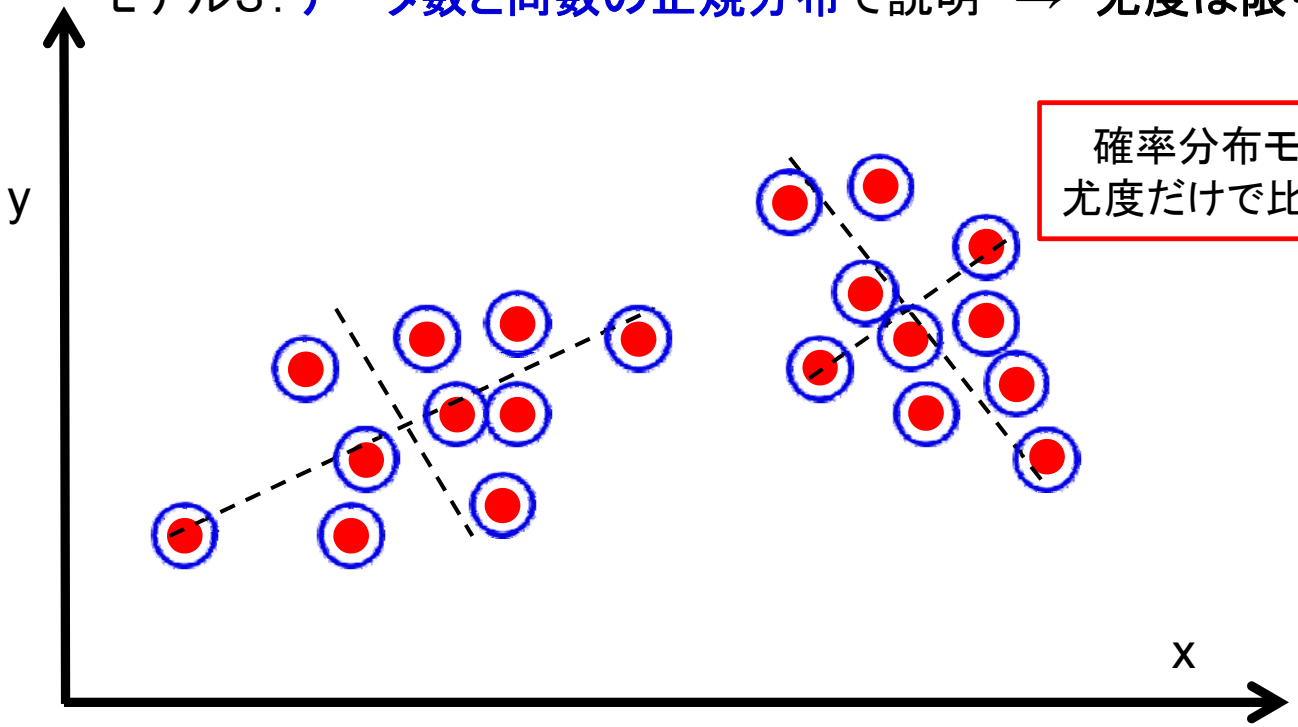
最尤推定と確率モデル構築：　究極の「データ解析」

膨大な事例データから、結果を説明する**単純なルール**を見つける

モデルから観測データが発生する確率

モデル３：**データ数と同数の正規分布**で説明　→　**尤度は限りなく大きくできる**

確率分布モデルの妥当性は
尤度だけで比べても意味がない



最尤推定と確率モデル構築：　究極の「データ解析」

- データを説明するモデルの評価法：**この評価値が小さいほど良いモデル**

AIC　赤池情報量基準（Akaike's Information Criterion）

$$AIC = -2\ln L + 2k \Rightarrow \text{最小化}$$

$-2\ln L$：モデルの尤度　　$2k$：モデルの大きさ

ここでLは最大尤度、kは自由パラメータ数

MDL 最小記述長（Minimum Description Length）

$$MDL = -\ln L + \frac{k\ln n}{2} \Rightarrow \text{最小化}$$

$-\ln L$：モデルの尤度　　$\frac{k\ln n}{2}$：モデルの大きさ

ここでLは最大尤度、n はデータ数、kは自由パラメータ数

まとめ

（1） 情報量（平均情報量）
（2） エントロピー
（3） データ圧縮・符号化・量子化
（4） 決定木（分類木）による知識表現・決定木の生成方法・評価
（5） 確率モデルの構築と評価

【参考文献】

今井秀樹： 情報理論、昭晃堂（昭和59年）

麻生英樹、津田宏治、村田昇： パターン認識と学習の統計学、岩波書店（2003年）

Stuart Russell and Peter Norvig: Artificial Intelligence – A Modern Approach, Pearson Education Inc. (2003)