情報処理基礎

新井康平

1

授業計画

- 9/28：イントロ
- 10/5：統計の意味
- 10/12：Excelの機能
- 10/19：Excelの機能
- 10/26：情報量の計算
- 11/9：データの相関
- 11/16：相関分析
- 11/23：回帰分析
- 11/30：確率密度関数
- 12/7：二項分布、正規分布
- 12/14：χ^2分布
- 12/21：指数分布
- 1/11：仮説検定
- 1/18：仮説検定
- 1/25：仮説検定
- 2/1：試験

2

Excel

3

確率の基礎

ある事象が起きる確率を p とするとき、それが起きたという情報が持つ情報量 S は、

$$S = -log(p) \tag{1.1}$$

である。場合分けの数が 2 のべき数でないときの情報量も与えられる。たとえば，サイコロを投げたときに出た目が 5 であるという情報量は、その確率は 1/6(場合分けの数が 6) なので、

$$S = -log(1/6) = 2.58 \tag{1.2}$$

となる。また、トランプのカードを一枚引いてそのカードがエースであるとき、

$$S = -log(1/13) = 3.70 \tag{1.3}$$

4

「UFO キャッチャ」の縦・横の位置を決めて縫いぐるみ人形を取るゲームにおける情報を考えてみよう。この場合、縫いぐるみのおいてある位置の縦・横両方のの情報が必要である。判り易く縦・横に 16 の目盛りが刻んであったとすると、升目は全部で 256 個あり、その中から求める縫いぐるみのおいてある位置を特定するために 256 の升目から一つを選ばなければならない。まず、縦の升目を決めると、後は 16 の升目から一つを決めれば良いことになる。すなわち、256 通りあた決め方が 16 通りに減ったのである。更に、横の升目を決めると、一つの升目、縫いぐるみのある升目が決定できることになる。すると、最初から 256 の升目から一つを特定するに要する情報量は、縦の升目を選ぶために必要な情報量と、横の升目を決めるために必要な情報量の和になっているはずである。これを情報量の相加の法則と呼ぶ。

上述の 2 つの例に合致する関数として、対数関数が最もふさわしい。すなわち、種類の数、場合分けの数 N の対数で情報量 I とするものである。

$$I = \log(N) \tag{1.4}$$

対数の底の取り方によって情報量の単位が異なる。すなわち 2 であればビット、自然数 e であればナット、10 であればハートレイである（情報量の定義に対数関数を導入したのは R. V. Hartley (1928) である）。1 ハートレイは 3.32 ビット、1 ナット

5

$$\log_a x = \frac{\log_b x}{\log_b a} \tag{1.5}$$

である。数字の場合、種類数は 10 であるので、

$$I = \log_2 10 = 3.3[bit] \tag{1.6}$$

以下同様に、英字の場合、$\log_2 26 = 4.7[bit]$、仮名の場合は、$\log_2 45 = 5.5[bit]$ となる。常用漢字は、$\log_2 21945 = 10.925[bit]$ であり、人名、地名漢字を加えても、数字の情報量のたかだか 4 倍である。種類数および場合分けの数の逆数は確率になるので、確率: P と情報量: I の関係は以下のようになる。

$$I = -\log P \tag{1.7}$$

6

- 対数関数の計算
- 乱数生成

演習 1

さいころなげの事象の情報量をビットの単位で計算せよ。これを 1 から 6 までの整数の乱数を発生するプログラムを作り、試行回数を変えて、それぞれの目のでる確率を計算し、情報量に直せ。

演習 2

あるクラスに n 人の生徒がいる。少なくともそのうちの 2 人が同じ誕生日である確率を計算せよ。但し、1 年 365 日とし、実際に n を仮定して、確率を計算せよ。

7

各事象 a_i を元とする事象の集合 $A = \{a_1, a_2, ..., a_n\}$ を全事象、そのうちのいくつかの集合 (たとえば $B = \{a_1\}$ 等) を部分集合という。さいころの例であれば、1 から 6 の目のでる事象が全事象であり、そのうち、奇数の目のでる事象の集合が部分集合である。また、和集合、共通集合の定義もでき、たとえば、奇数の目のでる事象の集合 $C = \{a_1, a_3, a_5\}$ と偶数の目のでる集合 $D = \{a_2, a_4, a_6\}$ との和集合は全事象になるとか、

$$C\{a_1, a_3, a_5\} \cup D\{a_2, a_4, a_6\} = A\{a_1, ..., a_6\} \tag{1.9}$$

$$p(A) = p(C \cup D) = p(C) + p(D) = 1 \tag{1.10}$$

$$p(C) = 1/2, p(D) = 1/2 \tag{1.11}$$

または、両集合の共通集合は空集合 $\emptyset$ であるとか定義できる。

$$C\{a_1, a_3, a_5\} \cap D\{a_2, a_4, a_6\} = \emptyset \tag{1.12}$$

$$p(\emptyset) = 0 \tag{1.13}$$

また、この例のように共通集合を持たない事象どうし (偶数の目がでる事象と奇数の目のでる事象のような場合) を排反事象という。

8

演習 3

1 組のトランプ (52 枚) から、エースを取り出す確率、ダイヤを取り出す確率及びダイヤのエースを取り出す確率を計算せよ。

9

さて、さいころ投げの試行を繰り返すことを考え、i 番目が a_1 であって、$i+1$ 番目も a_1 である確率はどの様に表されるであろうか? これは、条件付き確率と呼ばれ、$p(a = a_1 \mid b = a_1)$ と表される。また、さいころふたつを同時に投げて、一方が a1 であって、他方も a1 である確率は、結合確率または同時確率と呼ばれ、$p(c = a_1, d = a_1)$ と表される。最初に a_1 が起きて、それが条件となって a_1 が起こる確率がこの結合確率 (a_1 と a_1 が同時に起こる確率) に等しいので、次式が成り立つ。

$$p(a = a_1)p(a = a_1 \mid a = a_1) = p(a = a_1, a = a_1) \tag{1.14}$$

$$p(a)p(b \mid a) = p(a, b) \tag{1.15}$$

また、$p(a, b) = p(b, a)$ であり、$p(b, a) = p(b)p(a \mid b)$ であるので、

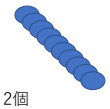
$$p(a \mid b) = \frac{p(a)p(b \mid a)}{p(b)} \tag{1.16}$$

10

演習 4

ある犬のまえに、10 個の容器の入った α、β の 2 つの部屋があり、α、β の 10 個の容器には、それぞれ、2 及び 3 個のみが餌が入っているものとし、それ以外は空であるとする。さいころを降って、1、2 の目がでれば、α から、それ以外の目がでれば、β の部屋から、それぞれ一つの容器を取り出すものとする。このとき、餌にありつく確率を求めよ。

- α :1,2
- 2個



β :3,4,5,6

3個

11

いま、カラオケボックスに a,b,c,d の 4 人が入り、歌を唄うとして、誰が唄っているのかについて「レジ」に 2 つの符号で通報するものとする。このとき、4 人のうち a から d の順に歌が好きで、長時間唄いたがり、表 1.4 のような確率でマイクを握っているものとする。また、通報する符号は、表 1.4 の右に示されているような 2 種類の符号の組合せであるとする。

表 1.1 歌を唄っている確率

人物	確率	符号
a さん	1/2	00
b さん	1/4	01
c さん	1/8	10
d さん	1/8	11

12

最初の符号を受けて、それが"0"であると、a さんか b さんに唄っている人は限られ、それぞれ符号を受ける前に 1/2、1/4 であった確率は,1/2/(1/2+ 1/4)=2/3 および 1/2/(1/2+1/4)=1/3 に増加する。次に"1 "の符号を受けると、b さんの唄っている確率は、1/3 から 1 に増加し、a さんの唄っている確率は 0 になる。このように、歌を唄っている人に関する符号、または、情報を得ることによって、それに関する曖昧さが少なくなって行くことがわかる。最初の符号を x, 最後の符号を y とし、最初の符号が運んだ b さんに関する情報量を $I(b, x)$ とすると、これは、b さんの事前確率 $p(b)$ と、$p(b \mid x)$ の関数として与えられると考えられるので、

$$I(b, x) = f[p(b), p(b \mid x)] \tag{1.17}$$

と表せる。次に最後の符号によって運ばれる情報量 $I(b, y \mid x)$ は、

$$I(b, y \mid x) = f[p(b \mid x), p(b \mid x, y)] \tag{1.18}$$

となる。x, y を一つの符号と考え、これを同時に受け取る場合の情報量は、

$$I(b, xy) = f[p(b), p(b \mid xy)] \tag{1.19}$$

13

これは $I(b, x)$ と $I(b, y \mid x)$ の和になっているはずである。

$$I(b, xy) = I(b, x) + I(b, y \mid x) \tag{1.20}$$

これを情報量相加の法則と呼ぶが、これが成り立つ関数: f のひとつに対数関数があり、情報量をこの関数で定義する理由のひとつになっている。

$$I(b, x) = \log p(b \mid x)/p(b) \tag{1.21}$$

$$I(b, y \mid x) = \log p(b \mid xy)/p(b \mid x) \tag{1.22}$$

$$I(b, xy) = \log p(b \mid xy)/p(b) \tag{1.23}$$

14

演習 5

チケット売り場の窓口にお客が来ている確率が 5/8、お客のいない確率が 3/8 であるとする。お客が来ている時は赤ランプが店頭につき、来ていない時には青ランプがつくと言う。ところが、受付がこのランプの点滅を行っており、お客が来ている時、赤のランプをつける的中率が 3/4、お客がいない時に青のランプをつける的中率が 1/2 であるとする。この時、お客のいるかいないかに関する事象と、ランプをつけるか否かに関する事象の相互情報量を求めよ。

15

エントロピー

16

ある事象 A が起こる確率を p とすれば、起こらない確率は $q = 1 - p$ である。このとき、A が起こったとすると、その情報量は $-logp$ であり、起こらなかったとすると、その情報量は $-logp$ である。そのため、全体の情報量の期待値は、

$$H = -log(p) - log(q) \tag{1.24}$$

となる。この H をエントロピー (平均情報量) と呼ぶ。これを一般化して、確率事象が 1 から n 個あり、それらの生起確率を $p_1, ..., p_n$ とすると、そのエントロピーは、

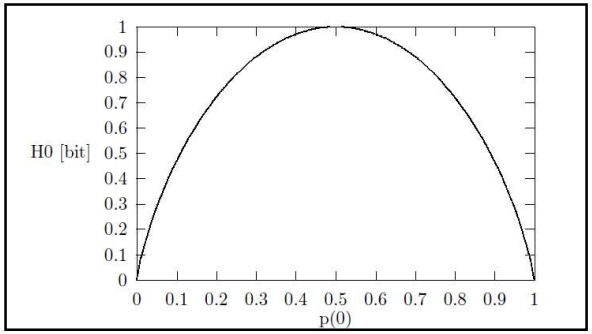
$$H = -\sum_{i=1}^{n} log(p_i) \tag{1.25}$$

と表せる。エントロピーは増大する。たとえば、消費者が全然知識のない商品 A と B があり、その商品の販売比率を $p, q(= 1 - p)$ とすると、価格やデザインなどに特別の違いがなければ、$p = q = 0.5$ になる。その理由は、このエントロピーが、

$$H = -plog(p) - (1 - p)log(1 - p) \tag{1.26}$$

となるからであり、$p = q = 0.5$ のときに最大になるからである。

17



18

消費者が価格以外の知識のない商品 A と B があるとする。A の価格は a 円、B の価格は b 円である。この場合、商品 A と B の販売比率 p および $q = 1 - p$ は、費用の期待値とエントロピーの観点から以下のように求められる。消費者は、費用を最小にしようと行動する。費用の期待値は、$C = ap + bq$ である。また、消費者は価格以外の商品に関する情報を得ていないので以下のエントロピーが最大になるように商品を選択する。

$$H = -plog(p) - qlog(q) \tag{1.27}$$

これら費用の期待値を最小にするように、また、エントロピーを最大にするように消費者は商品選択を行う。すなわち、H/C が最大となるように p, q が決定すると考えられる。したがって、

$$x^a + x^b = 1 \tag{1.28}$$

の根を求め、$p = x^a, q = x^b$ を求める。

19

の根を求め、$p = x^a, q = x^b$ を求める。たとえば、$a = 1, b = 2$、すなわち、B 商品は A 商品の倍の価格であるとすると、

$$x + x^2 = 1$$

$$x = 0.618$$

$$p = x = 0.618$$

$$q = x^2 = 0.382 \tag{1.29}$$

となる。このとき、エントロピーは、

$$H = -0.618log(0.618) - 0.382log(0.382) = 0.959 \tag{1.30}$$

である。また、費用の期待値は、

$$C = ap + bq = 1 \times 0.618 + 2 \times 0.382 = 1.382 \tag{1.31}$$

となる。この価格に関する商品情報が与えられていない場合は、エントロピーが最大になる条件のみにて消費者は行動するので $H = -plog(p) - qlog(q) \rightarrow$ 最大となるように商品選択を行うと、$p = 0.5. a = 0.5. H = 1$ となって商品 $A. B$ ともに同数売れることになる。しかし、B の商品価格が A の倍であるという情報を与えた場合は 61.8%は A 商品を、また、38.2%は B 商品を購入することになる。

20