応用数学

秋葉洋哉

2024 年 6 月 23 日

1　線形代数

1.1　行列

ある連立方程式が与えられたとき、その連立方程式を行列で表すことができる。例えば、以下の連立方程式を考える。

$$5x_1 + 2x_2 = 10 \tag{1}$$
$$3x_1 + 4x_2 = 11 \tag{2}$$

行列は連立方程式の各係数を縦横に並べたものであり、上述した連立方程式は行列を用いて以下のように表される。

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{3}$$
$$\begin{pmatrix} 5 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 11 \end{pmatrix} \tag{4}$$

この時、$\mathbf{A}$ のことを行列と呼ぶ。また、$\mathbf{x}$ のことをベクトルと呼ぶ。

1.2 逆行列

行列 $\mathbf{A}$ に対して、行列 $\mathbf{B}$ が存在して以下の式が成り立つとき、行列 $\mathbf{B}$ を行列 $\mathbf{A}$ の逆行列と呼ぶ。

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I} \tag{5}$$

ただし、$\mathbf{I}$ は単位行列である。逆行列の求め方は、掃き出し法、余因子法、逆行列の公式などがある。ここで、逆行列の存在しない行列を考える。連立方程式の解が無く、一組に決まらない場合、逆行列は存在しない。たとえば、以下の連立方程式を考える。

$$x_1 + 2x_2 = 3 \tag{6}$$
$$2x_1 + 4x_2 = 6 \tag{7}$$

行列で表すと以下のようになる。

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \end{pmatrix} \tag{8}$$

この時、行列 $\mathbf{A}$ の逆行列は存在しない。このことを端的に表すと、2 本のベクトルが同一直線上にある場合、逆行列は存在しないということを示している。つまり、

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \vec{v}_1 \\ \vec{v}_2 \end{pmatrix} \tag{9}$$

において、二つのベクトルに囲まれる平行四辺形の面積が 0 の場合、逆行列は存在しないということである。この「面積」を行列式と呼び、以下のように表す。

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = \begin{vmatrix} \vec{v}_1 \\ \vec{v}_2 \end{vmatrix} \tag{10}$$

行列式は判別式 (determinant) とも呼ぶ。なぜなら、行列式が 0 の場合、連立方程式の解が一意に定まらないため、解が存在しないということを示すからである。行列式は、計算上 4 つの重要な性質を有する。

1. 同じ行ベクトルが含まれていると行列式はゼロになる。
2. 1 つのベクトルがλ倍されると行列式はλ倍される。
3. 他の成分がすべて同じで、i 番目のベクトルだけが違った場合、行列式の足し合わせになる。
4. 行を入れ替えると符号が変わる。

行列式の計算方法は、余因子展開、行列の対角成分を用いる方法、行列の固有値を用いる方法などがある。2×2 の行列式の計算方法は、以下のようになる。

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \tag{11}$$

3×3 の行列式の計算方法は、以下のようになる。

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh \tag{12}$$

1.3 固有値と固有ベクトル

正方行列 $\mathbf{A}$ に対して、以下の式が成り立つとき、λ を行列 $\mathbf{A}$ の固有値、$\mathbf{v}$ を行列 $\mathbf{A}$ の固有ベクトルと呼ぶ。

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{13}$$

固有値と固有ベクトルは、行列 $\mathbf{A}$ の固有方程式を解くことで求めることができる。固有方程式は以下のように表される。

$$\begin{vmatrix} a-\lambda & b \\ c & d-\lambda \end{vmatrix} = 0 \tag{14}$$

固有値と固有ベクトルは、行列 $\mathbf{A}$ の対角化に用いられる。行列 $\mathbf{A}$ が対角化可能であるとは、行列 $\mathbf{A}$ が固有ベクトルを基底とする行列 $\mathbf{P}$ と対角行列 $\mathbf{D}$ を用いて以下のように表されることをいう。

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1} \tag{15}$$

ただし、$\mathbf{P}$ は固有ベクトルを列ベクトルとして並べた行列、$\mathbf{D}$ は固有値を対角成分に持つ行列である。

固有値と固有ベクトルは、行列の対角化に用いられるだけでなく、主成分分析、特異値分解、クラスタリングなどの機械学習アルゴリズムにも広く用いられる。

1.4 特異値分解

$\mathbf{A}$ が正方行列でない場合、特異値分解という方法で、正方行列と同様に分解することができる。行列 $\mathbf{A}$ に対して、以下の式が成り立つとき、$\mathbf{U}$ を左特異ベクトル、$\mathbf{V}$ を右特異ベクトル、$\mathbf{\Sigma}$ を特異値と呼ぶ。

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{16}$$

特異値の求め方は、以下の通りである。まず、特異値が存在するには、2 つの単位ベクトル $\vec{v}, \vec{u}$ が以下の条件を満たす必要がある。

$$\mathbf{A}\vec{v} = \sigma\vec{u} \tag{17}$$
$$\mathbf{A}^T\vec{u} = \sigma\vec{v} \tag{18}$$

このような特殊な単位ベクトルがある時、特異値分解できる。ここで、σ を Σ とし、行列で表すと、以下のようになる。

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \end{pmatrix} \tag{19}$$

さらに、$\vec{v}, \vec{u}$ を列ベクトルとして並べた行列を $\mathbf{V}, \mathbf{U}$ とし、式 (17) と式 (18) を行列で表すと、以下のようになる。

$$\mathbf{A}\mathbf{V} = \mathbf{U}\mathbf{\Sigma} \tag{20}$$
$$\mathbf{A}^T\mathbf{U} = \mathbf{V}\mathbf{\Sigma}^T \tag{21}$$

これらの式を変形すると、

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{-1} \tag{22}$$
$$\mathbf{A}^T = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^{-1} \tag{23}$$

となり、これらの積は、

$$\mathbf{A}\mathbf{A}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{-1}\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^{-1} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^{-1} \tag{24}$$

となる。これより、$\mathbf{A}\mathbf{A}^T$ を固有値分解することで、特異ベクトルと特異値の 2 乗が求められる。特異値分解は、主成分分析、画像圧縮、レコメンデーションシステムなどに広く用いられる。

■1 章 参考文献

2　統計学

2.1　集合

S という集合に属する要素を a とすると、a について以下のように表す。

$$a \in S \tag{25}$$

要素が a 以外にも、b, c, d がある場合、S について以下のように表す。

$$S = \{a, b, c, d\} \tag{26}$$

また、集合の内部に M という部分集合がある場合、以下のように表す。

$$M \subset S \tag{27}$$

確率・統計に登場する「事象」は、「集合」と同じ概念であると考えることができる。たとえば、サイコロを振るとき、出る目が「1」である事象を A とすると、以下のように表す。

$$A = \{1\} \tag{28}$$

また、サイコロを振るとき、出る目が「偶数」である事象を B とすると、以下のように表す。

$$B = \{2, 4, 6\} \tag{29}$$

このように決めておくことで、確率を数式で表すことができる。

和集合・共通部分・絶対補・相対補は記号を用いて以下で表す。

$$A \cup B : A\text{と}B\text{の和集合} \tag{30}$$
$$A \cap B : A\text{と}B\text{の共通部分} \tag{31}$$
$$U \ A = \bar{A} : A\text{の絶対補} \tag{32}$$
$$B \ A : A\text{の相対補} \tag{33}$$

これらの概念も、確率・統計において重要な役割を果たす。

2.2　確率

確率は大きく分けると「頻度確率」と「ベイズ確率」の 2 つに分類される。

頻度確率は、試行を繰り返し、その結果を観測することで確率を求める方法である。例えば、10 本のくじを引いて当選する確率を調べたところ、100 回試行したときに 10 回当選したという事実から、当選確率は 10% であるという結論を導くことができる。

ベイズ確率は、事前確率と事後確率を用いて確率を求める方法である。例えば、医者が患者に対して、「あなたは 40% の確率でインフルエンザに感染しています」と診断したとしたとき、その医者は、別の医者 100 人が同じ診断を行った結果、40 人が同様に感染していると診断するだろう、と考えているのではなく、患者の症状や病歴などを考慮して、事前確率を修正することで、事後確率を求める。

頻度確率は客観確率とも呼ばれ、ベイズ確率は主観確率とも呼ばれる。

確率を定義する。ある事象 A が起こる確率を $P(A)$ とすると、以下のように表す。

$$P(A) = \frac{n(A)}{n(U)} \tag{34}$$

ただし、$n(A)$ は事象 A が起こる回数、$n(U)$ は全事象が起こる回数である。

ここで、集合 A と集合 B の共通部分の確率を求める。

$$P(A \cup B) = P(A)P(B|A) \tag{35}$$

ただし、$P(B|A)$ は、条件付き確率と呼ばれる。

2.3 条件付き確率

条件付き確率は、ある事象 B が与えられた下で、別の事象 A が起こる確率のことであり、

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{36}$$

で表される。例えば、雨が降っている条件下で交通事故に遭う確率を求める場合、雨が降っているという事象 B が起こったときに、雨が降っているかつ交通事故に遭うという事象 $A \cap B$ が起こる確率を求めることで導出することができる。確率で表しているが、実際には、事象 $A \cap B$ が起こる回数を事象 B が起こる回数で割ることで求めることができる。つまり、

$$P(A|B) = \frac{n(A \cap B)}{n(B)} \tag{37}$$

であるといえる。ただし、お互いの事象が独立である場合、

$$P(A|B) = P(A) \tag{38}$$

であるから、

$$P(A \cap B) = P(A)P(B) \tag{39}$$

となることに注意する。

2.4 ベイズ則

ベイズ則は、条件付き確率を用いて、事象 A が起こったときに、事象 B が起こる確率を求める方法である。ベイズ則は以下のように表される。

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{40}$$

ここで、具体的な数値を用いてベイズ則の例を示す。例えば、ある病気の検査があり、その検査の結果が陽性であった場合、実際にその病気にかかっている確率を求めることを考える。

病気である、病気でない、検査で陽性、検査で陰性という 4 つの事象を考え、それぞれの事象を $D, \bar{D}, T, \bar{T}$ とする。また、以下の情報が与えられているとする。

- 病気にかかっている人の割合： $P(D) = 0.01$
- 病気にかかっている人が検査で陽性となる確率： $P(T|D) = 0.99$
- 病気にかかっていない人が検査で陽性となる確率： $P(T|\bar{D}) = 0.03$

このとき、ベイズ則を用いて実際に病気にかかっている確率を求めることができる。まず、分子である、病気にかかっている人が検査で陽性となる確率を求める。

$$P(T|D)P(D) = 0.99 \times 0.01 = 0.0099 \tag{41}$$

次に、分母である、検査で陽性となる確率を求める。

$$\begin{aligned} P(T) &= P(T|D)P(D) + P(T|\bar{D})P(\bar{D}) && (42)\\ &= 0.99 \times 0.01 + 0.03 \times 0.99 = 0.0399 && (43) \end{aligned}$$

最後に、ベイズ則を用いて、病気にかかっている確率を求める。

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T)} && (44)\\ &= \frac{0.0099}{0.0399} && (45)\\ &\approx 0.247 && (46) \end{aligned}$$

したがって、検査で陽性となった場合、実際に病気にかかっている確率は約 24.7% であるという結論を導くことができる。

2.5　記述統計と推測統計

統計学は、大きく分けて「記述統計」と「推測統計」の 2 つに分類される。

記述統計は、データを集計し、要約することで、データの特徴を把握する方法である。例えば、平均値、中央値、最頻値、分散、標準偏差などがある。

推測統計は、データから得られた結果をもとに、未知の母集団の特性を推測する方法である。例えば、母集団の平均値、中央値、最頻値、分散、標準偏差などを推測する。推測統計では、母集団から標本を抽出し、その標本から得られた結果をもとに、母集団の特性を推測することで、未知の母集団の特性を推測することができる。

2.6 記述統計その 1 確率変数と確率分布

確率変数は、ある事象に対して確率を割り当てる変数のことであり、確率分布は、確率変数が取りうる値とその確率の関係を示す関数のことである。例えば、サイコロを振るとき、出る目を確率変数 X とし、その確率分布を以下のように表す。

$$P(X=1)=\frac{1}{6} \tag{47}$$

$$P(X=2)=\frac{1}{6} \tag{48}$$

$$P(X=3)=\frac{1}{6} \tag{49}$$

$$P(X=4)=\frac{1}{6} \tag{50}$$

$$P(X=5)=\frac{1}{6} \tag{51}$$

$$P(X=6)=\frac{1}{6} \tag{52}$$

このように、確率変数が取りうる値とその確率の関係を示す関数を確率分布と呼ぶ。

確率分布には、ベルヌーイ分布、マルチヌーイ (カテゴリカル) 分布、二項分布、正規 (ガウス) 分布、ポアソン分布、指数分布、ガンマ分布、ベータ分布、ディリクレ分布などが存在する。

ベルヌーイ分布

ベルヌーイ分布は、コイントスを 1 回行うイメージである。ひしゃげたコインを投げたとき、表が出る確率を μ、裏が出る確率を $1-\mu$ とし (普通のコインなら $\mu=1/2$ のはずだ)、x を 1 のとき成功、0 の時失敗を表すパラメータとすると、ベルヌーイ分布は以下のように表す。

$$P(x|\mu)=\mu^x(1-\mu)^{1-x} \tag{53}$$

この時、$E[x]=\mu, Var(x)=\mu(1-\mu)$ である。

二項分布

二項分布は、ベルヌーイ分布のコイントスを複数回行ったときのイメージである。コインを n 回投げたとき、表が出る回数を x 回とすると、二項分布は以下のように表す。

$$P(x|\mu,n)=\frac{n!}{x!(n-x)!}\mu^x(1-\mu)^{n-x} \tag{54}$$

この時、$E[x]=n\mu, Var(x)=n\mu(1-\mu)$ である。

マルチヌーイ (カテゴリカル) 分布

マルチヌーイ (カテゴリカル) 分布は、サイコロを振るイメージである。サイコロを振ったとき、出る目が 1 の確率を p_1、出る目が 2 の確率を p_2、出る目が 3 の確率を p_3、出る目が 4 の確率を p_4、出る目が 5 の確率を p_5、出る目が 6 の確率を p_6 とすると、マルチヌーイ (カテゴリカル) 分布は以下のように表す。

$$P(x|\mu)=\prod_{k=1}^{K}\mu_k^{x_k} \tag{55}$$

正規 (ガウス) 分布

正規 (ガウス) 分布は、ベル型の連続分布であり、多くの自然現象に適用される。正規 (ガウス) 分布は以下のように表す。

$$P(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{56}$$

正規 (ガウス) 分布は、真の分布が分からなくても、サンプルが多ければ、中心極限定理により正規分布に近づくという性質がある。

2.7 記述統計その 2 期待値

事象を $X_1, X_2, \ldots X_n$ とし、それぞれの確率を $p(X_1), p(X_2), \ldots p(X_n)$、確率変数の値を $f(X_1), f(X_2), \ldots f(X_n)$ とすると、期待値は以下のように表す。

$$E[f(X)] = \sum_{k=1}^{n} p(X = x_k) f(X = x_k) \tag{57}$$

$$E[f(X)] = \int p(X = x) f(X = x) dX \tag{58}$$

2.8 記述統計その 3 分散

分散は、確率変数の散らばり具合を示す指標であり、以下のように表す。

$$Var(X) = E[(f_{(X=x)} - E_{(f)})^2] \tag{59}$$
$$= (\text{二乗の平均}) - (\text{平均の二乗}) \tag{60}$$
$$= E[X^2] - E[X]^2 \tag{61}$$

共分散は、2 つの確率変数の関係性を示す指標であり、以下のように表す。

$$Cov(X, Y) = E[(f_{(X=x)} - E_{(f)})(g_{(Y=y)} - E_{(g)})] \tag{62}$$
$$= E[XY] - E[X]E[Y] \tag{63}$$

分散を見ることで、平均値からのばらつき具合を知ることができ、共分散を見ることで、2 つの確率変数の関係性を知ることができる。例えば、f と g の動き方が似ている時、共分散は正の値を取り、f と g の動き方が逆の時、共分散は負の値を取る。

分散では、元のデータの単位が 2 乗されるため、元のデータと同じ単位で表すためには、標準偏差を用いる。標準偏差は、分散の平方根であり、以下のように表す。

$$\sigma(X) = \sqrt{Var(X)} \tag{64}$$

2.9 推測統計その 1 推定

推測統計では、推定と検定の 2 つの方法がある。中でも推定は、母集団の特性を推測する方法であり、大きく分けて「点推定」と「区間推定」の 2 つに分類される。

点推定は、母集団の特性を 1 つの数値で推定する方法であり、平均値、中央値、最頻値などを求める推定方法である。区間推定は、母集団の特性を区間で推定する方法であり、信頼区間、予測区間などを求める推定方法である。ここでは、点推定で求められた推定値の例として標本平均 (母集団から取り出した標本の平均値) の特性を挙げる。

1. 一致性：サンプルサイズが大きくなれば、標本平均は母平均に近づく。
2. 不偏性：サンプル数がいくらであっても、その期待値は母集団の値と同様である。

推定の際には、推定量と推定値の違いを理解することが重要である。推定量は、パラメータを推定するために利用する数値の計算方法や計算式のことであり、推定値は、推定量を用いて計算された具体的な数値のことである。例えば、2 次関数を微分して出た導関数は推定量であり、傾きは推定値である。

サンプルサイズとサンプル数

何度標本の抽出を行ったか、という回数をサンプル数と呼び、1 回の抽出で何個体を調べたか、という数をサンプルサイズと呼ぶ。例えば、各都道府県から無作為に 1000 人を抽出して平均身長を算出したとき、サンプル数は 47、サンプルサイズは 1000 となる。

2.10 推測統計その 2 標本分散

標本分散は、母集団の分散を推定するために用いられる統計量であり、以下のように表す。

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{65}$$

ここで、σ^2 は標本分散、n はサンプルサイズ、X_i は各データ、$\bar{X}$ は標本平均である。標本分散は、一致性を満たす一方で、不偏性を満たさない。これは、たくさんのデータのばらつき具合と、少数のデータのばらつき具合を考えた時、少数のデータのばらつきの方が小さくなると考えることができるためである。

母集団の分散を推定する際には、不偏分散を用いることが多い。不偏分散は、標本分散をサンプルサイズで補正したものであり、以下のように表す。

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{66}$$

ここで、s^2 は不偏分散である。

3　情報科学

3.1　導入

情報の量とは何だろうか。抽象的な情報量という概念を具体的に表すために、以下の例を考えてみる。左の

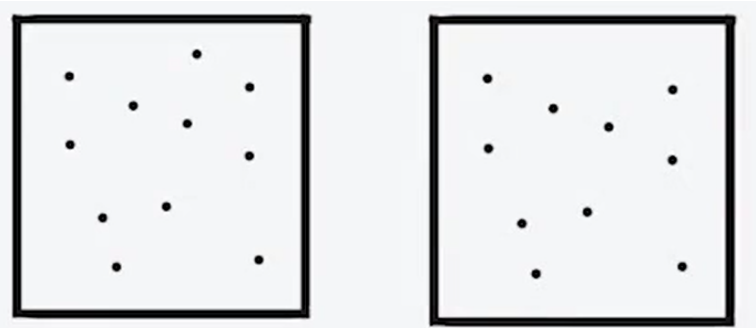

図 1　箱の中の点の増分 1

箱には 11 個の点があり、右の箱には 10 個の点がある。この差分は 1 個である。同様に以下の図をみてみる。

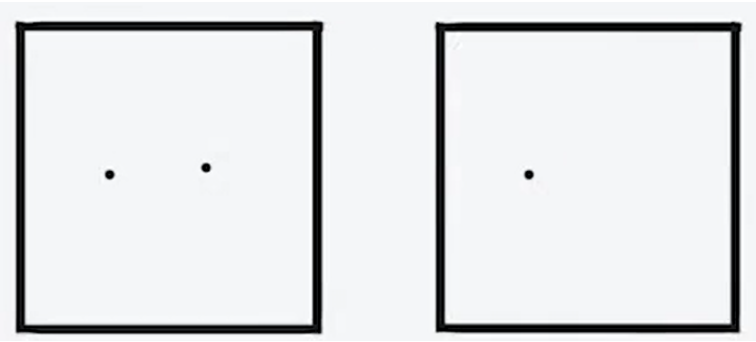

図 2　箱の中の点の増分 2

左の箱には 2 個の点があり、右の箱には 1 個の点がある。この差分は 1 個である。図 1 と図 2 を比較すると、どちらも差分が 1 個だが、明らかに 2 個から 1 個になった場合の方が、情報の増え方が大きく感じる。この経験則から、w を変化後の点の数、δw を変化前との点の差分とすると、情報の増え方を表す指標として、以下の「比率」を考えることができる。

$$\frac{\delta w_1}{w_1} = \frac{1}{10} \tag{67}$$

$$\frac{\delta w_2}{w_2} = \frac{1}{1} \tag{68}$$

この比率は、情報の増え方を表すものであるから、この式を積分することで、情報量を表す指標を導出することができる。

$$\int \frac{1}{w} dw = \log w \tag{69}$$

この時、w は事象の数であるため、確率 p に置き換えると、以下のように表すことができる。

$$I(p) = -\log p \tag{70}$$

この式は、「自己情報量」と呼ばれる。対数の底が 2 のとき、単位はビット (bit) となり、底がネイピア数のとき、単位はナット (nat) となる。この式は、情報の珍しさ (=確率の小ささ) を定量化したものであり、確率が小さいほど、情報量が大きくなることを示している。

3.2 シャノンエントロピー

自己情報量の期待値を取ることで、シャノンエントロピー (単純にエントロピーともいう) を定義することができる。

$$\begin{aligned} H(x) &= E(I(x)) && (71)\\ &= -E(log(P(x))) \quad (\log(\mathrm{P(x)}) \text{ が確率変数になる}) && (72)\\ &= -\sum p(x)\log p(x) \quad (\text{期待値は、確率} \times \text{確率変数の総和}) && (73) \end{aligned}$$

シャノンエントロピーは、新しく得られる情報の量を表す指標と考えることができる。例えば、表が出る確率が 0.99 の極端にひしゃげたコインのコイントスを行う場合、そのコインを投げた時に新しく得られる情報の量 (つまり裏が出る事象数) はとても小さく、シャノンエントロピーも小さな値となる。このことを図 3 で確認するには、x 軸が 1.0 付近の y 軸の値を読み取ればよい。他に、例えば表が出る確率が 0.25, 裏が出る確率を 0.75 とすると、シャノンエントロピーは

$$-0.25\log 0.25 - 0.75\log 0.75 \approx 0.811 \tag{74}$$

となる。

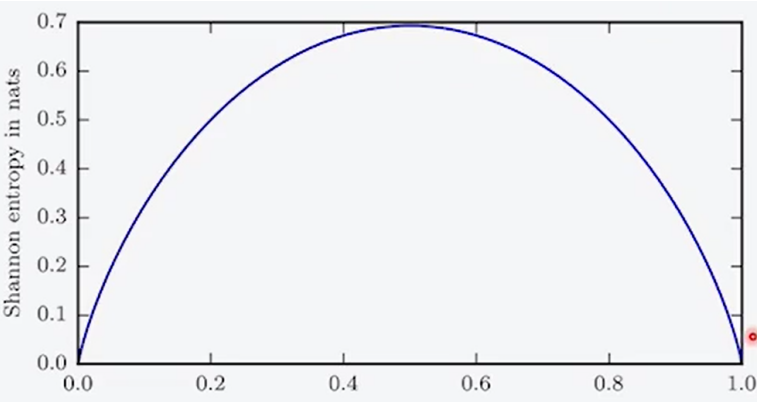

図 3　ひしゃげたコインのコイントスを行う場合のシャノンエントロピー

3.3 相互情報量

以下では細工されたサイコロの出る目の確率分布表を示す。このサイコロは、1 が出る確率が 0.5, それ以外が出る確率が 0.1 となるような細工がされている。このサイコロを振った時、例えば 4 の目が出た時に得られ

細工されたサイコロの出る目の確率分布表						
確率変数 X	1	2	3	4	5	6
確率 P	0.5	0.1	0.1	0.1	0.1	0.1

図 4　細工されたサイコロの出る目の確率分布

る情報量は、$-\log_2 0.1$ である。また、サイコロを振り出た目を表す確率変数を X_1 とすると、エントロピー $H(X_1)$ は、$-0.5\log_2(0.5) - 0.1\log_2(0.1) * 5$ で表される。

次にもう一度サイコロを振り、出た目を表す確率変数を X_2 とする。$X_1 = X_2$ のとき 1, $X_1 \neq X_2$ のとき 0 となる確率変数 Y を定義する。この時、2 つのエントロピー $H(X_1)$ と $H(Y)$ の和はどのような値をとるだろうか。この値は結合エントロピーで表される。結合エントロピーは、2 つの確率変数のエントロピーの和で定義される。

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y) \tag{75}$$

$$\tag{76}$$

表 1 を参照にすると、上記のサイコロの例での X_1 と Y の結合エントロピー $H(X_1,Y)$ は以下のように表すことができる。

表 1　結合エントロピーの計算

	Y=0	Y=1
X=1	0.5*0.1*5	0.5*0.1
X=2	0.1-0.1*0.1	0.1*0.1
X=3	0.1-0.1*0.1	0.1*0.1
X=4	0.1-0.1*0.1	0.1*0.1
X=5	0.1-0.1*0.1	0.1*0.1
X=6	0.1-0.1*0.1	0.1*0.1

次に条件付きエントロピーを導入する。条件付きエントロピーは、ある確率変数が与えられたときのエントロピーであり、以下のように表すことができる。

$$H(X|Y) = -\sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(y)} \tag{77}$$

この式は、X が与えられたときの Y のエントロピーを表す。

相互情報量は、もともとのエントロピー $H(X)$ と条件付きエントロピー $H(X|Y)$ の差であるため、上記サイコロの例における相互情報量を計算すると、以下のように表すことができる。

$$I(X_1;Y) = H(X_1) - H(X_1,Y) \tag{78}$$

$$\tag{79}$$

3.4 カルバック・ライブラーダイバージェンス

カルバック・ライブラーダイバージェンスは、2 つの確率分布の違いを表す指標であり、$Q(x)$ を最初に見積もった確率、$P(x)$ を真の確率とすると、以下のように表す。

$$D_{KL}(P||Q) = \mathbb{E}_{x\sim P}[\log \frac{P(x)}{Q(x)}] \tag{80}$$

$$= \mathbb{E}_{x\sim P}[\log P(x) - \log Q(x)] \tag{81}$$

ただし、$\mathbb{E}$ は引数 $[\log P(x)/Q(x)]$ を平均したものを表す。この引数部分を自己情報量を用いて変形すると、以下のように表すことができる。

$$\log P(x)/Q(x) = I(Q(x)) - I(P(x)) \tag{82}$$

$I(Q(x))$ が最初に分かっていた情報量、$I(P(x))$ が新しく得られた情報量であるため、カルバック・ライブラーダイバージェンスは、Q という状態から P という状態に変化したときに、どのくらい新しい情報が得られたかを表す指標であるといえる。ただし、Q から P に変化した場合と、P から Q に変化した場合で、カルバック・ライブラーダイバージェンスの値は異なり、距離とは異なることに注意する。カルバック・ライブラーダイバージェンスは、以下で変形することができる。

$$D_{KL}(P||Q) = \sum P(x)(\log P(x) - \log Q(x)) \tag{83}$$

$$= \sum P(x) \log \frac{P(x)}{Q(x)} \tag{84}$$

3.5 交差エントロピー

KL ダイバージェンスの一部分を取り出したものが交差エントロピーである。Q についての自己情報量を P の分布で平均している。

$$H(P,Q) = -\sum P(x)\log Q(x) \tag{85}$$

$$= -\sum P(x)\log\frac{Q(x)}{P(x)} + \sum P(x)\log P(x) \tag{86}$$

$$= D_{KL}(P||Q) + H(P) \qquad = -\mathbb{E}_{x\sim P}[\log Q(x)] \tag{87}$$

交差エントロピーは分類問題における誤差関数としてよく用いられ、真の確率分布 ($P(x)$) と推定した確率分布 ($Q(x)$) に対して、2 つの確率分布の誤差を定量的に求めるために使われる。

■2 章 参考文献

1. 病気に罹患している確率は? ベイズの定理で求める方法を解説『Python で動かして学ぶ!』シリーズ (https://codezine.jp/article/detail/14581)
2. うさぎでもわかる情報量・エントロピー・相互情報量（情報理論）(https://www.momoyama-usagi.com/entry/info-entropy)