

第 1 章

数学的準備

ここでは本稿の目的である手書き数字の認識を行うニューラルネットワークの構築に必要な数学的事項を説明する。大まかに分けて、線形代数、確率論、情報理論の知識が必要になる。

1.1 線形代数

1.1.1 スカラー、ベクトル、行列、テンソル

スカラー (scalar) は -1.4 , 1 , π などの単一の数のことをいう。本稿の数式上では a, b, c, \dots のようにアルファベットのイタリック体で書き表すことにする。

スカラーを 1 次元状に並べたものをベクトル (vector) という。本稿ではベクトルを \boldsymbol{x} のように、ボールドイタリック体の小文字のアルファベットで表す。ベクトルの成分は x_1, y_i のようにスカラーに要素番号を添えることで表す。

2 次元の数の配列を行列 (matrix, pl. matrices) という。本稿では A などのイタリック体の大文字で行列を表し、 A_{ij} のように 2 つの添字を付すことで成分を表すことにする。

3 次元以上の数の配列をテンソル (tensor) といい、本稿では A のようにサンセリフ体で表記する。テンソル A の (i, j, k) 成分は A_{ijk} のように表す。

1.1.2 積の計算

行列 A について、その成分 A_{ij} と A_{ji} を入れ替えた行列を転置行列 (transposed matrix) または単に転置 (transpose) といい、 A^\top と表記する。すなわち $A_{ij} = (A^\top)_{ji}$ 。ベクトルは 1 つの列ベクトルからなる行列とみなせば、ベクトルの転置は行ベクトルである。転置の例を次に示す。

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \longrightarrow A^\top = \begin{bmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \end{bmatrix},$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \longrightarrow \boldsymbol{x}^\top = [x_1, x_2, \dots, x_n].$$

スカラー a については、それを 1×1 の行列とみなせば転置をとっても変わらない。すなわち $a^\top = a$ である。

$m \times n$ の行列 A と $n \times p$ の行列 B に対して、行列積 (matrix product) $C = AB$ が次のように定義できる。

$$C_{ij} = \sum_k A_{ik} B_{kj}. \quad (1.1)$$

この行列 C のサイズは $m \times p$ になる。

2つの同じサイズのベクトル $\mathbf{x} = [x_1, \dots, x_n]^\top$ と $\mathbf{y} = [y_1, \dots, y_n]^\top$ について、ドット積 (dot product) は行列積 $\mathbf{x}^\top \mathbf{y}$ となり、これはスカラーである。

$$\mathbf{x}^\top \mathbf{y} = [x_1, \dots, x_n] \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = x_1 y_1 + \dots + x_n y_n = \sum_{k=1}^n x_k y_k. \quad (1.2)$$

$\mathbf{x}^\top \mathbf{y}$ はスカラーであるから、 $(\mathbf{x}^\top \mathbf{y})^\top = \mathbf{x}^\top \mathbf{y}$ である。また式 (1.2) から分かるように、 $\mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x}$ である。ゆえに $\mathbf{x}^\top \mathbf{y} = (\mathbf{x}^\top \mathbf{y})^\top = \mathbf{y}^\top \mathbf{x}$ である。

行列のアダマール積 (Hadamard product) を計算することがある。同じサイズの行列 A と B について、そのアダマール積 $A \odot B$ は、それぞれの行列の成分ごとの積を計算した行列である。すなわち $(A \odot B)_{ij} = A_{ij} \cdot B_{ij}$ である。この定義はより一般にテンソルにも拡張できる。

数式の見た目をシンプルにするために、アインシュタインの縮約記法 (Einstein summation convention) を使うことがある。テンソルの成分についての計算時に、同じ項の中で同じ添字が現れた場合、総和の記号 \sum を省略して書くのだ。この記法を用いれば、式 (1.1) は $C_{ij} = A_{ik} B_{kj}$ 、式 (1.2) は $\mathbf{x}^\top \mathbf{y} = x_k y_k$ と書ける。アインシュタインの縮約記法を用いるときは必ずその旨を表すようにする。

1.1.3 行列の微分

行列 A に対して、その A_{ij} の成分を偏微分演算子 $\partial/\partial A_{ij}$ に置き換えた行列を ∇_A と書く。もちろんこれは演算子であるから、何らかの関数に左から作用させたときに実体を持つ。

$$(\nabla_A f)_{ij} = \frac{\partial f}{\partial A_{ij}}.$$

1.1.4 ノルム

ベクトル $\mathbf{x} = [x_1, \dots, x_n]^\top$ の L^p ノルム (L^p norm) を

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$$

と定義する。

L^2 ノルムは数学的に扱いやすく、よく用いられるため、ユークリッドノルム (Euclidean norm) ともよばれる。特に L^2 ノルムの2乗は、そのベクトル自身とのドット積に一致する。

$$\|\mathbf{x}\|_2^2 = x_1^2 + \dots + x_n^2 = \mathbf{x}^\top \mathbf{x}.$$

1.2 確率論

1.2.1 確率論を使うことの意義

コンピュータ科学ではほとんどの場合、完全に決定論的で、確実な対象を扱うが、機械学習では不確実・非決定論的な量を扱わなければならない。不確実性を生み出す要因は3つある。1つめは扱う対象そのものが確率性を持つ場合（例：量子力学や完璧にシャッフルされたカードなど）、2つめは扱う対象の振る舞いを決める変数を観測できない・しない場合（例：モンティ・ホール問題において回答者の回答が当たりかハズレかは決定論的に決まっているが、ドアを開けるまで結果は不確実である）、最後は観測した情報を破棄した場合（例：物体の位置を離散化したとき、正確な位置は不確実となる）である。

確実だが複雑な規則よりも、不確実だが単純な規則を用いるほうが実用的であるため、確率論が用いられる。

頻度確率（frequentist probability）とは、反復可能な試行を無限回くりかえしたときに、ある事象が現れる頻度という意味での確率のことである。いっぽう反復可能な試行でなくとも、信念の度合い（degree of belief）、可能性の大きさという意味での確率をベイズ確率（Bayesian probability）という。確率論ではこれら2つの意味での確率を、同等に扱う。

確率論はある命題のもっともらしさが与えられたときに、対象の命題が真であるもっともらしさを決定するための、不確実性を扱うための論理学の拡張とみなすことができる。

1.2.2 確率変数と確率分布

確率変数（random variable）とは起こりうる事象に対応した、ランダムな値を取れる変数のことである。確率変数を x のようにローマン体の小文字で書き、 x がとりうる値は x のようにイタリック体で書く。確率変数がベクトル値のときはそれらは太字で \mathbf{x} , \mathbf{x} のように書く。確率変数は離散値でも連続値でもよい。たとえばサイコロを振ったときの「6の目が出る」という事象は $x = 6$ に対応させる、コイントスをしたときの「裏が出る」「表が出る」という事象をそれぞれ $x = 0$, $x = 1$ に対応させる。このように確率変数は、事象の集合（標本空間）から数の集合への関数と考えることもできる。

確率質量関数（probability mass function, PMF）とは、離散型確率変数に対してその値をとるときの確率に対応させた関数のことであり、大文字 P で書き表す。たとえば確率 p で表（ $x = 1$ ）、確率 $1 - p$ で裏（ $x = 0$ ）が出るようなコインでコイントスをするとき、確率質量関数は

$$P(x = x) = \begin{cases} p, & (x = 1) \\ 1 - p & (x = 0) \end{cases}$$

とかける（これはベルヌーイ分布とよばれる）。このコイントスを n 回行ったうち k 回だけ表が出る確率は

$$P(y = k) = \binom{n}{k} p^k (1 - p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k}$$

となる（これは二項分布とよばれる）。このとき「確率変数 y は母数 (n, p) の二項分布 $B(n, p)$ に従う」といい、これを $y \sim B(n, p)$ と表記する。

$x = x$ かつ $y = y$ であるときの確率 $P(x = x, y = y) = P(x, y)$ は同時確率分布とよばれる。

関数 P が離散型確率変数 x の確率質量関数であるためには以下の性質を満たさなければならない。

- 定義域は確率変数のとりうる値すべての集合である。
- その集合を A とすると、 $\forall x \in A$ に対して $0 \leq P(x) \leq 1$.
- 正規化されている (normalized) : $\sum_{x \in A} P(x) = 1$.

たとえば、確率変数 x が k 個の異なる離散値 x_1, x_2, \dots, x_k をとるとき、確率質量関数を

$$P(x = x_i) = \frac{1}{k}$$

とすれば一様分布を定義することができる。すべての i で和をとれば

$$\sum_{i=1}^k P(x = x_i) = \frac{1}{k} \sum_{i=1}^k 1 = \frac{1}{k} \cdot k = 1$$

となって、正規化されていることがわかる。

離散型確率変数に対して確率質量関数とよんでいたものを、連続型確率変数に対しては確率密度関数 (probability density function, PDF) とよぶ。確率密度関数 p は以下の性質を満たさなければならない。

- 定義域は確率変数のとりうる値すべての集合である。
- その集合を I とすると、 $\forall x \in I$ に対して $0 \leq p(x)$. ただし $p(x) \leq 1$ である必要はない。
- $\int_I p(x) dx = 1$.

$x = x$ であるときの確率は $p(x)$ と直接得られるわけではなく、代わりに x が微小区間 $[x, x + \delta x]$ の値をとるときの確率が $p(x)\delta x$ で与えられる。区間 $[a, b]$ に x が存在する確率は $\int_{[a,b]} p(x) dx$ で求められる。確率密度関数の例として、一様分布がある。

$$u(x; a, b) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & x \notin [a, b]. \end{cases}$$

積分すれば1になる。 x が区間 $[a, b]$ において一様分布にしたがうことを、 $x \sim U(a, b)$ と表す。

1.2.3 条件付き確率と確率の連鎖律

ある事象が起こったもとでの別の事象が起こる確率を条件付き確率 (conditional probability) という。 $x = x$ が与えられたもとでの $y = y$ が起こる条件付き確率 $P(y = y | x = x) = P(y | x)$ は

$$P(y = y | x = x) = \frac{P(y = y, x = x)}{P(x = x)} \quad (1.3)$$

と定義される。

多数の確率変数 x_1, \dots, x_n に対する同時確率分布は式 (1.3) を用いれば、1つの確率変数についての条件付き確率に分解してそれらの積として表現できる。このことを確率の連鎖律 (chain rule of probability) という。

$$P(x_1, \dots, x_n) = P(x_1) \prod_{i=2}^n P(x_i | x_1, \dots, x_{i-1}).$$

1.2.4 ベイズの定理

条件付き確率 $P(y | x)$ を知っているときに、別の条件付き確率 $P(x | y)$ を求めたい状況はよくある。確率分布 $P(x)$ が分かっているならば、これは次のベイズの定理 (Bayes' theorem) を用いることで計算できる。

$$P(x | y) = \frac{P(x)P(y | x)}{P(y)} = \frac{P(x)P(y | x)}{\sum_x P(x)P(y | x)}. \quad (1.4)$$

1.2.5 期待値と分散

確率変数 x が確率分布 P に従うとき、関数 $f(x)$ の平均を期待値 (expectation) といい、記号 $\mathbb{E}_{x \sim P}[f(x)]$ と表す。ただし確率変数 x が確率分布 P に従うことが明らかな場合は、単に $\mathbb{E}[f(x)]$ と書く。離散型確率変数についての期待値の定義は

$$\mathbb{E}_{x \sim P}[f(x)] = \sum_x P(x)f(x)$$

であり、連続型確率変数の期待値は積分を用いて

$$\mathbb{E}_{x \sim p}[f(x)] = \int p(x)f(x) dx$$

と定義される。

また関数 $f(x)$ のばらつき度合いを表すのに分散 (variance) が用いられる。

$$\text{Var}[f(x)] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[(f(x))^2] - (\mathbb{E}[f(x)])^2.$$

1.3 情報理論

1.3.1 自己情報量とシャノン・エントロピー

「地球は自転している」という情報よりも「明日地球に隕石が衝突する」という情報の方が意外性が高く、より大きな価値をもつ気がする。このような直感的な「情報の価値」のようなものを数学的に表現したい。具体的には

- 起こりやすい事象の情報量は少なく、確実に起こる事象の情報量はないとする
- 起こるのが珍しい事象ほど情報量は大きい
- 独立な事象については情報量は足し算で表される。例えば「コイントスを2回行ったところ表が2回出た」という事象の情報は、「コイントスを1回行ったところ表が出た」という事象のそれよりも、2倍の大きさをもつ

というようなものを数式で表現したい。この3つの性質を満たすためには次のような量を作ればよい。

$$I(x) = -\log P(x).$$

これを事象 $x = x$ の自己情報量 (self-information) という。本稿では \log は常に自然対数を表すものとし、その場合の自己情報量の単位はナット (nats) という。

自己情報量は1つの事象のみについての情報量であるが、全体の情報量の平均をシャノン・エントロピー (Shannon entropy) という。

$$H(x) = H(P) = \mathbb{E}_{x \sim P}[I(x)] = - \sum_i P(x_i) \log P(x_i).$$

たとえば確率 p で表が出るコインでコイントスをするときのシャノン・エントロピーは $H(P) = -p \log p - (1-p) \log (1-p)$ となる。この関数を図示すると図 1.1 のようになり、 $p = 0, 1$ (結果が確実にわかっている) ときに $H(P) = 0$, $p = 1/2$ (結果は不確実) のときに最大値を取ることがわかる。一般に離散型確率変数に対する確率分布のシャノン・エントロピーが最大になるのは一様分布のときである。

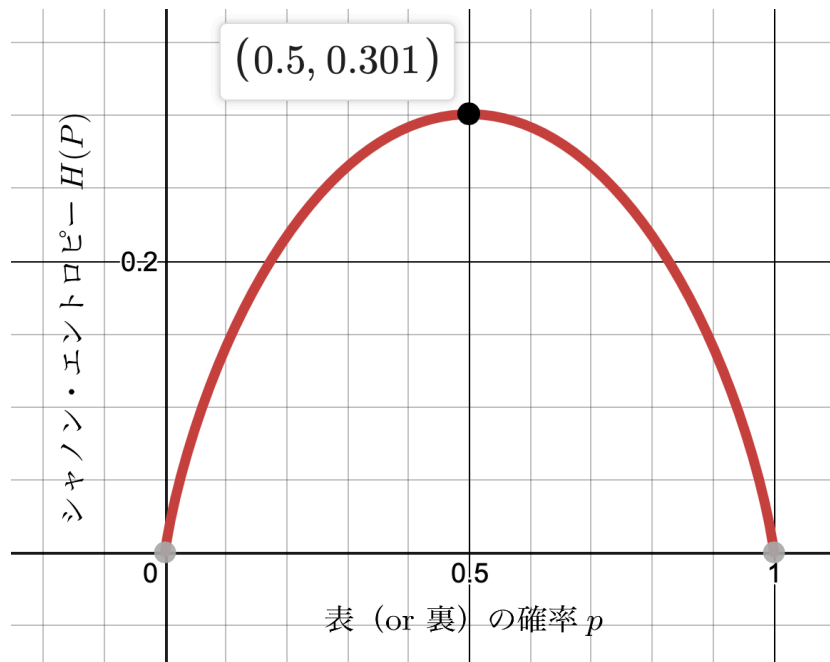


図 1.1 二値確率変数のシャノン・エントロピー

1.3.2 KL ダイバージェンスと交差エントロピー

同じ確率変数 x に対して異なる確率分布 $P(x)$ と $Q(x)$ があつたとき、それらがどれほど異なるのかを表す量として KL ダイバージェンス (Kullback-Leibler divergence) がある。

$$D_{\text{KL}}(P\|Q) = \mathbb{E}_{x \sim P} \left[\log \frac{P(x)}{Q(x)} \right] = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}.$$

KL ダイバージェンスは非負の量であり、等号成立は P と Q が同じときである。よってこれは P と Q の距離のような概念と考えられるが、一般に $D_{\text{KL}}(P\|Q) \neq D_{\text{KL}}(Q\|P)$ であるため距離の公理の1つ $d(x, y) = d(y, x)$ を満たさず、距離と呼ぶのは正しくない。

また次の量を交差エントロピー (cross-entropy) という。

$$H(P, Q) = H(P) + D_{\text{KL}}(P\|Q).$$

これを少し式変形をすると

$$\begin{aligned} H(P, Q) &= H(P) + D_{\text{KL}}(P \| Q) \\ &= - \sum_i P(x_i) \log P(x_i) + \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)} \\ &= - \sum_i P(x_i) \log Q(x_i) \end{aligned}$$

となり、 $\mathbf{x} \sim P$ のもとでの $Q(x)$ のシャノン・エントロピー $\mathbb{E}_{\mathbf{x} \sim P}[-\log Q(x)]$ を表しているとわかる。KL ダイバージェンスと比較して取り除かれている部分 $\sum_i P(x_i) \log P(x_i)$ は Q に依存しないため、交差エントロピーを Q に関して最小化することは KL ダイバージェンスを最小化することと等価である。交差エントロピーはのちにコスト関数としてよく用いる。