船津・小寺研究室 新人向け事前課題

井上 貴央

2019/03/02

当研究室では主に, 機械学習の手法を利用して研究を進めています. 機械学習の手法の理解には, 基本的な微分積分学の知識や線形代数学, 確率論, 統計学といった数学の知識が欠かせません. これらについては研究室に入ってからの新人研修でも折に触れて復習しますが, 春休み中にも簡単に復習をしていただきたいと思いこのようなテキストを作りました.

以下の**問題**を解いたものを, pdf 形式で **3/22 (金)** までに井上 (`t_inoue@chemsys.t.u-tokyo.ac.jp`) まで提出してください (**問題**は全部で 11 問あります). **発展問題**は少し難しめなので解く必要はありませんが, 余力があればやってみてください. 本テキストには問題以外に, 復習用の解説と問題を解くための**ヒント**をつけています. 必要であれば確認してください. 答案は TEX や Word 等で作成されているのが望ましいですが, もちろん手書きのスキャン/写真などを pdf にまとめていただいても構いません. また, 内容について質問があれば対応しますのでいつでも聞いてください.

なお, 機械学習の分野で扱うのはふつうは実数値のデータです. このため, **以下では基本的に変数は実数値をとるものとして扱ってください**.

1 微分積分学の復習

機械学習の手法の理解に必要な微分積分学の基礎知識は, それほど多くありません[*1]. 基本的な微積分の演算ができて, 簡単な最適化問題が解ければひとまず十分かと思いますので, 問題演習を通して復習しましょう.

1.1 微積分演算

まずは, ウォーミングアップとして 1 変数関数の微分を復習しましょう. 積の微分, 商の微分, 逆関数の微分, 合成関数の微分など, いろいろと公式がありましたが, 覚えているでしょうか. **問題 1** を解いて計算をチェックしてみましょう.

[*1] もちろん機械学習の理論をしっかりやるには相応の知識が必要ですが.

問題 1 (1 変数関数の微分)

以下の関数の導関数を求めてください.

(1) $f_1(x) = x \log x \quad (x > 0)$.
(2) $f_2(x) = \exp\left(-\frac{1}{2}(x-m)^2\right) \quad (m \in \mathbb{R}$ は定数).
(3) $f_3(x) = \arctan x$.
(4) $f_4(x) = \frac{1}{1+\exp(-ax)} \quad (a \in \mathbb{R}$ は定数).

ヒント

(1) 積の微分公式を使います.
(2) 合成関数の微分公式を使います.
(3) 逆関数の微分公式を使います.
(4) 商の微分公式と合成関数の微分公式を使います. ちなみに, この関数は**シグモイド関数** (sigmoid function) と呼ばれています.

多変数関数の**偏微分** (partial derivative) は機械学習をやる上で多用されます. 偏微分する変数以外を定数だと思って微分すればよかったですね. **問題 2** で練習してみましょう.

問題 2 (多変数関数の偏微分)

以下の関数の x_1 に関する偏導関数を求めてください.

(1) $f_1(x_1, x_2) = x_1 \cos x_2$.
(2) $f_2(x_1, x_2) = \frac{2x_1x_2}{x_2^2 - x_1^2}$.
(3) $f_3(x_1, x_2, x_3) = \sqrt{x_1^2 + x_2^2 + x_3^2}$.
(4) $f_4(x_1, x_2) = \arctan \frac{x_2}{x_1} \quad (x_1 \neq 0)$.

ヒント

変数 x_2 を定数とみなして x_1 で微分すればよいです. 合成関数の微分の際に計算ミスをしないように注意しましょう.

微分ほど頻繁ではありませんが, 積分の計算も出てくることがあります[*2]. 特に, 次の**発展問題 1** の定積分は **Gauss 積分**と呼ばれる有名な定積分で, 正規分布がらみの計算でよく出てきます.

[*2] 例えば, 平均, 分散の計算や Bayes 推論の際に利用することがあります.

発展問題 1 (Gauss 積分)

Gauss 積分について

$$\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{\lambda}{2}(x-\mu)^2}\,\mathrm{d}x = \sqrt{\frac{2\pi}{\lambda}} \quad (\mu \in \mathbb{R},\ \lambda > 0 \text{ は定数})$$

を証明してください.

ヒント

変数変換 $t = \sqrt{\frac{\lambda}{2}}(x-\mu)$ により,

$$\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t$$

という形の定積分が出てきます. この計算方法はたくさんありますが, 有名なものとしては

$$\left(\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t\right)^2$$

を二重積分の形に変形し, 極座標変換を用いて計算する方法があります. 多重積分の座標変換の際には, **ヤコビアン** (Jacobian) の絶対値の項が現れることに注意しましょう.

1.2 最適化問題

機械学習では, 関数 $f(\boldsymbol{x})$ の値を最小/最大にする変数 $\boldsymbol{x}$ を見つける, という操作が頻繁に行われます. この操作を関数 f の**最適化** (optimization) といい, 最適化問題を解く際に微分がよく利用されます. 関数 $f(\boldsymbol{x})$ の最大化問題は, 関数 $-f(\boldsymbol{x})$ の最小化問題と同じですので, 以下では最小化問題を扱います. まずは, 最適化で利用される用語を**定義 1.1** に整理しておきましょう.

定義 1.1 (最適化問題)

集合 S は $\mathbb{R}^n$ の部分集合とします. 変数 $\boldsymbol{x}$ を S の中で動かして関数 $f(\boldsymbol{x})$ を最小化する問題を次のように表記します ("s.t." は "subject to" の略です):

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{x} \in S. \end{aligned}$$

最適化する対象の関数 f を**目的関数** (objective function) といい, 集合 S を**実行可能領域** (feasible region) といいます. 実行可能領域 S を決定する条件式を**制約条件** (constraint) といい, S の代わりに制約条件が与えられることも多いです.

目的関数の最小値を**大域的最適値** (global optimal value), 大域的最適値を与える変数を**大域的最適解** (global optimal solution) といいます. 普通, 単に「最適値」とか「最適解」とか言う場合は, 大域的最適値や大域的最適解のことを指します. これに対して, 目的関数の極小値

を**局所的最適値** (local optimal value), 極小値を与える極小解を**局所的最適解** (local optimal solution) とも呼ばれます. □

まずは, 機械学習でよく現れる**凸関数** (convex function) の最適化をやってみることにしましょう. 凸関数とは, 単純に言ってしまえば「下に出っ張っている関数」のことです[*3]. 制約条件がない状況で凸関数を最小化するときは, 極小解を見つければそれがそのまま最小解になることが知られています[*4].

1 階微分可能な n 変数関数 $f(\boldsymbol{x})$ $(\boldsymbol{x} \in \mathbb{R}^n)$ の極小解が $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ であるためには, $\bar{\boldsymbol{x}}$ は f の**停留点** (stationary point), つまり $\bar{\boldsymbol{x}}$ での勾配 $\nabla f(\bar{\boldsymbol{x}})$ が

$$\nabla f(\bar{\boldsymbol{x}}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\bar{\boldsymbol{x}}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\bar{\boldsymbol{x}}) \end{pmatrix} = \boldsymbol{0}$$

をみたすことが必要なのでした[*5]. これは必要条件なので一般には逆が成り立たないのですが[*6], 実は凸関数では逆も成り立ち, 必要十分条件になります.

以上の事実を利用して, 次の**問題 3** を解いてみてください.

問題 3 (凸関数の制約なし最適化)

以下の目的関数の最小解とそのときの最小値を求めてください. なお, その際に以下の関数が凸関数になっていることを用いても構いません.

(1) $f_1(x) = ax^2 + bx + c \quad (a, b, c \in \mathbb{R}$ は定数, $a > 0)$.
(2) $f_2(x) = x \log x \quad (x > 0)$.
(3) $f_3(x_1, x_2, x_3) = 4x_1^2 + 4x_2^2 + 4x_3^2 - 2x_1x_2 - 2x_2x_3 + 2x_3x_1$.
(4) $f_4(x) = |x - a| \quad (a \in \mathbb{R}$ は定数$)$.

ヒント

(1)〜(3) の目的関数は凸関数なので, 停留点を求めればそれが最小解です. (3) では, 連立方程式を解く必要がありますが, 線形代数の知識を利用すると少々楽かもしれません. なお, (4) のように微分不可能な点がある場合はこの方法は使えませんが, 関数の形状が容易に分かるので問題ありません.

[*3] 正確には, 関数 $f(\boldsymbol{x})$ が凸関数であるとは, 任意の $\boldsymbol{x}, \boldsymbol{y}$ と任意の $\lambda \in [0, 1]$ に対して, $f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$ をみたすことをいいます.

[*4] 詳細は省きますが, 制約条件がある場合でも, その条件が「良い条件」であれば極小解が最小解になります.

[*5] 高校のときにも「ある点で極値をとるならば, その点での微分係数が 0」というのをやったかと思います. この主張はその多変数版です.

[*6] 1 変数版の場合でも反例として関数 $f(x) = x^3$ の $x = 0$ での様子がよく挙げられていたかと思います. 点 $x = 0$ での微分係数は $f'(0) = 0$ ですが, f は $x = 0$ で極値を取りません (変曲点になっている).

制約条件として等式制約がついている場合は, **Lagrange の未定乗数法** (method of Lagrange multiplier) を用います[*7]. 一般には Lagrange の未定乗数法の条件も必要条件ですが, 目的関数が凸関数で等式制約がすべて 1 次式になっている問題[*8]では, Lagrange の未定乗数法の条件は必要十分条件になります. この事実を利用すると, 次の**発展問題 2** が解けます.

発展問題 2 (Lagrange の未定乗数法)
等式制約 $x_1 + x_2 + x_3 = 1$ のもとで, 関数 $f(x_1, x_2, x_3) = 4x_1^2 + 4x_2^2 + 4x_3^2 - 2x_1x_2 - 2x_2x_3 + 2x_3x_1$ の最小解 x_1, x_2, x_3 とそのときの最小値を求めてください. なお, 目的関数が凸関数になっていることを利用しても構いません.

ヒント
目的関数が凸関数で等式制約が 1 次式なので, Lagrange の未定乗数法の条件が最小解であるための必要十分条件になります. よって, Lagrange の未定乗数法の条件を書き出し, その条件を満たすものが最小解になります.

目的関数が凸関数でない場合の最小化問題では, 停留点が極小値を与える保証はありませんので, きちんと停留点が極小値を与えることを確認しなければなりません. このためには, **Hesse 行列** (Hessian matrix) の正定値性を確認する必要があり, 少々面倒になります[*9].

発展問題 3 (一般の関数の制約なし最適化)
関数 $f(x_1, x_2) = x_1^4 + x_2^4 - (x_1 + x_2)^2$ の最小解 x_1, x_2 とそのときの最小値を求めてください.

ヒント
まずは, 停留点をすべて求め, その中から Hesse 行列が正定値になる極小解を求めましょう. 極小解がすべて求まった後はその中で最小のものを見つければ, それが最小解です.

[*7] ちなみに, 制約条件に不等式制約がついている場合は, Lagrange の未定乗数法の一般化である **Karush–Kuhn–Tucker 条件** (KKT 条件) を使います.

[*8] **凸計画問題** (convex programming) と呼ばれる問題の一種です.

[*9] 目的関数が 1 変数関数の場合, 極小解付近では 1 階微分係数が単調増加しており, 極小解での 2 階微分係数が正になっていました. 多変数関数で 2 階微分係数に対応するのが Hesse 行列であり, 2 階微分係数が正であることに対応するのが Hesse 行列の正定値性です.

2　線形代数学の復習

機械学習では, 線形代数の知識がよく利用されます. 特に, 行列を用いた種々の計算や, 内積, ノルム, 固有値/固有ベクトルといったものは, 機械学習の手法を勉強する上で多く用いられています. これらについて復習をしてみましょう. なお, 単にベクトルと書いた場合, ふつう縦ベクトルを指しているものとします.

2.1　行列の計算演習

行列の和/差や行列の積の計算は覚えているでしょうか. 行列の和/差については, 同じ成分同士を足したり引いたりすればよかったですね. 行列の積については少々特殊な計算をするのでした. 以下に定義をまとめます.

定義 2.1 (行列の和/差と積)
行列 $A = (a_{ij})$, $B = (b_{ij})$ が $m \times n$ 行列で形が同じであるとき, これらの和/差 $A \pm B$ が定義できます. このとき行列 $A \pm B$ も $m \times n$ 行列で, その (i, j)-成分は

$$(A \pm B)_{ij} = a_{ij} \pm b_{ij}$$

で計算されます.

行列 $A = (a_{ij})$ が $\ell \times m$ 行列で, 行列 $B = (b_{ij})$ が $m \times n$ 行列となっていて, A の横ベクトルの次元と B の縦ベクトルの次元が等しいときに積 AB を定義できます. このとき積 AB は $\ell \times n$ 行列で, その (i, j)-成分は

$$(AB)_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$$

となります. □

実数の掛け算と違って, 行列の積は可換でないことに注意しましょう. つまり, 行列 A, B について積 AB と BA が定義できるとき, 一般には $AB \neq BA$ です (もちろん, $AB = BA$ が成立することもあります). また, $\ell \times m$ 行列 $A = (a_{ij})$ と m 次元ベクトル $\boldsymbol{b} \in \mathbb{R}^m$ の積も, $\boldsymbol{b}$ を $m \times 1$ 行列とみなして計算すれば良いです.

また, 逆行列も重要な概念です.

定義 2.2 (逆行列, 正則)
行列 A は n 次正方行列とします. 行列 A に対し, n 次正方行列 X で $AX = XA = I_n$ (I_n は n 次単位行列) をみたす行列が存在するとき, この行列 X を A の**逆行列** (inverse) といい, A^{-1} で表します. 逆行列が存在する行列は**正則** (regular) であるといいます. □

逆行列の計算方法はどのようなものだったかを, 一度確認しておくと良いでしょう. 正則であるかどうかを確認するための条件としては, 以下の条件が使いやすいです.

定理 2.1 (正則性の同値条件)
行列 A は n 次正方行列とします. 行列 A が正則であるためには, A の行列式が $\det A \neq 0$ となることが必要十分です. □

また, 逆行列については次の性質も成り立ちます.

定理 2.2 (正則行列の積とその逆行列)
行列 A, B は n 次正方行列で正則であるとします. このとき, 行列 AB も正則で, その逆行列は $(AB)^{-1} = B^{-1}A^{-1}$ となります. □

もう一つよく出てくる演算として, 行列の転置があります.

定義 2.3 (転置)
行列 $A = (a_{ij})$ が $m \times n$ 行列のとき, A の**転置** (transpose) $A^\top$ は $n \times m$ 行列で, その (i, j)-成分は $(A^\top)_{ij} = a_{ji}$ となります. □

行列の転置についても, 次の性質を利用すると便利な場合があります.

定理 2.3 (行列の積とその転置)
行列 A は $\ell \times m$ 行列, 行列 B は $m \times n$ 行列であるとします. このとき, 行列 AB の転置行列は, $(AB)^\top = B^\top A^\top$ となります. □

以上の確認として, 実際に**問題 4** を解いて練習してみましょう.

問題 4 (行列演算)
行列 A, B, C が

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 2 & 0 & 2 \end{pmatrix}, \quad C = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix},$$

ベクトル $\boldsymbol{x}, \boldsymbol{y}$ が

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

となっています. 以下の計算をしてください.

(1) $(A^\top + 2B)\boldsymbol{y}$.
(2) $C^{-1}(C + I_3)C$.
(3) AB.
(4) $B^\top A^\top B^\top$.

(5) $\boldsymbol{x}^\top A^\top A\boldsymbol{x}$.

ヒント

(1) 転置をとって $A^\top$ を求めて計算します. 分配法則が成り立ちますので, それを使って計算しても構いません.

(2) C^{-1} の計算をしても良いですが, 実は必要ありません.

(3) 行列の積の計算方法に従って計算するだけです.

(4) 転置を取ってから積を計算するよりは, 転置の性質 $(X_1\cdots X_k)^\top = X_k^\top\cdots X_1^\top$ (行列の添字の順番に注意!) を上手く使って計算するのが楽です. また, 行列の積では結合法則 $(XY)Z = X(YZ)$ が成り立つので, (3) の結果が利用できます.

(5) $A^\top A$ を計算するのも一つの方法ですが, 実は $A\boldsymbol{x}$ の計算だけで十分です.

2.2 内積とノルム

ベクトルに対する演算として, 内積は重要です.

定義 2.4 (内積)

二つの n 次元ベクトル

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

に対して, それらの (標準) **内積** (inner product) を

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle := \boldsymbol{a}^\top \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i$$

と定義します[*10]. □

内積の性質としては, 次の双線形性が計算の際に便利です.

定理 2.4 (内積の双線形性)

定数 $c_1, c_2 \in \mathbb{R}$ と n 次元ベクトル $\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}, \boldsymbol{y}_1, \boldsymbol{y}_2$ に対して,

$$\langle c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2, \boldsymbol{y} \rangle = c_1 \langle \boldsymbol{x}_1, \boldsymbol{y} \rangle + c_2 \langle \boldsymbol{x}_2, \boldsymbol{y} \rangle$$
$$\langle \boldsymbol{x}, c_1\boldsymbol{y}_1 + c_2\boldsymbol{y}_2 \rangle = c_1 \langle \boldsymbol{x}, \boldsymbol{y}_1 \rangle + c_2 \langle \boldsymbol{x}, \boldsymbol{y}_2 \rangle$$

が成立します. □

[*10] 内積は標準内積以外にも存在しますが, ここでは深入りしません.

内積が定義されているとき, 内積を使ってベクトルのノルムを定義できます. ノルムはベクトルの「大きさ」を表す量であると解釈することができます.

定義 2.5 (ノルム, 正規化)

ベクトル $\boldsymbol{x}$ の (内積から誘導される) **ノルム** (norm) を

$$\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$

と定義します[*11]. 特にこのノルムは, 標準内積から誘導されるノルムで **Euclid ノルム**や **L2 ノルム**とも呼ばれます.

ベクトル $\boldsymbol{x}$ のノルムが $\|\boldsymbol{x}\| = 1$ となっているとき, ベクトル $\boldsymbol{x}$ は**正規化されている** (normalized) といいます. □

内積とノルムを用いて, 二つのベクトルのなす角を定義できます. これにより, 二つのベクトルの「方向」が似ているかどうかを議論できるようになります. とくに, 二つのベクトルの「方向」が本質的に異なることを表す「直交」という概念が定義できるようになる点が重要です.

定義 2.6 (2 ベクトルのなす角, 直交, 正規直交)

二つのベクトル $\boldsymbol{a}, \boldsymbol{b}$ の**なす角** (formed angle) θ $(0 \leq \theta \leq \pi)$ を

$$\cos\theta := \begin{cases} \frac{\langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|} & (\boldsymbol{a}, \boldsymbol{b} \neq \boldsymbol{0}) \\ 0 & (\boldsymbol{a} = \boldsymbol{0} \text{ または } \boldsymbol{b} = \boldsymbol{0}) \end{cases}$$

を満たす角と定義します. また, 二つのベクトル $\boldsymbol{a}, \boldsymbol{b}$ が**直交する** (orthogonal) とは, それらのなす角が $\frac{\pi}{2}$ であること, つまり $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = 0$ となることをいい, これを $\boldsymbol{a} \perp \boldsymbol{b}$ と書きます. ベクトルの集合 $\mathcal{V} = \{ \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \}$ に含まれる各ベクトルが正規化されており, かつ, どの二つのベクトルも互いに直交しているとき, すなわち

$$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{i,j} := \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

であるとき, $\mathcal{V}$ は**正規直交系** (orthonormal system) であるといいます[*12]. □

それでは, **問題 5** で内積とノルムの計算練習をしてみましょう.

問題 5 (内積とノルム)

ベクトル $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ が

$$\boldsymbol{x}_1 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{x}_3 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

[*11] 内積と同様に, やはりノルムにもいろいろあります.

[*12] ここで導入した記号 $\delta_{i,j}$ は **Kronecker のデルタ**と呼ばれます.

となっています. 以下の計算をしてください.

(1) $\ell_1 = \|\boldsymbol{x}_1\|$.
(2) $\boldsymbol{v}_1 = \ell_1^{-1}\boldsymbol{x}_1$.
(3) $\ell_2 = \|\boldsymbol{x}_2 - \langle \boldsymbol{x}_2, \boldsymbol{v}_1\rangle \boldsymbol{v}_1\|$.
(4) $\boldsymbol{v}_2 = \ell_2^{-1}(\boldsymbol{x}_2 - \langle \boldsymbol{x}_2, \boldsymbol{v}_1\rangle \boldsymbol{v}_1)$.
(5) $\ell_3 = \|\boldsymbol{x}_3 - \langle \boldsymbol{x}_3, \boldsymbol{v}_1\rangle \boldsymbol{v}_1 - \langle \boldsymbol{x}_3, \boldsymbol{v}_2\rangle \boldsymbol{v}_2\|$.
(6) $\boldsymbol{v}_3 = \ell_3^{-1}(\boldsymbol{x}_3 - \langle \boldsymbol{x}_3, \boldsymbol{v}_1\rangle \boldsymbol{v}_1 - \langle \boldsymbol{x}_3, \boldsymbol{v}_2\rangle \boldsymbol{v}_2)$.
(7) $\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle$, $\langle \boldsymbol{v}_2, \boldsymbol{v}_3\rangle$, $\langle \boldsymbol{v}_3, \boldsymbol{v}_1\rangle$.

ヒント
すべて, 内積とノルムの定義どおり計算するだけです. 計算では, 内積の双線形性 (定理 2.4) を使うのが楽です. 内積が実数値になることに注意しましょう. なお, この問題の背景には **Gram–Schmidt の正規直交化法**という, 一次独立な[*13]ベクトルから正規直交基底をつくるアルゴリズムがあります.

2.3 固有値/固有ベクトルと対角化

固有値と固有ベクトルは, 機械学習の手法のいたるところで現れる重要な概念です.

定義 2.7 (固有値/固有ベクトル)
行列 A は n 次正方行列とします. 行列 A に対して

$$A\boldsymbol{v} = \lambda\boldsymbol{v} \tag{1}$$

となる複素数 $\lambda \in \mathbb{C}$ とゼロベクトルでない n 次元ベクトル $\boldsymbol{v}$ をそれぞれ A の**固有値** (eigenvalue), **固有ベクトル** (eigenvector) といいます[*14]. □

式 (1) を変形すると,

$$(A - \lambda I_n)\boldsymbol{v} = \boldsymbol{0} \tag{2}$$

となります. もし, $A - \lambda I_n$ に逆行列が存在するとすると, $\boldsymbol{v} = \boldsymbol{0}$ となってしまいます. このため, ゼロベクトルでないベクトル $\boldsymbol{v}$ がこれを満たすためには, $A - \lambda I_n$ が正則でないこと, つまり

$$\det(A - \lambda I_n) = 0 \tag{3}$$

[*13] ベクトルの集合 $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ が一次独立であるとは,

$$\sum_{i=1}^{k} a_i \boldsymbol{v}_i = \boldsymbol{0} \;\Rightarrow\; a_i = 0 \quad (i = 1, \ldots, k)$$

が成立することを言います.
[*14] 固有値は複素数の範囲で考えることと, 固有ベクトルがゼロベクトルでないことに注意が必要です.

となることが必要十分です (定理 2.1). この方程式 (3) を**固有方程式** (characteristic equation) と呼びます. 固有方程式を解くことで, 固有値を求めることができます. 固有値が求まったら, これに対応する固有ベクトルを式 (2) から求めることができます.

もし, n 次正方行列 A の固有ベクトルが n 個あって, それらが一次独立になっているなら[*15], 行列の**対角化** (diagonalization) をすることができます. 行列 A の n 個の固有値を $\lambda_1, \ldots, \lambda_n$, それぞれに対応する固有ベクトルを $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ とすると,

$$A\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i \quad (i = 1, \ldots, n)$$

となります. これを一つの式でまとめて書くと,

$$A(\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n) = (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n) \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

となります. ここで, $(\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n)$ は縦ベクトル $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ が並んだ行列で, $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ は $\lambda_1, \ldots, \lambda_n$ が対角成分に並んだ対角行列です. 行列 P を $P = (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n)$ とすると, ベクトル $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ が一次独立であることから, P は正則で逆行列が存在します. つまり, P によって

$$P^{-1}AP = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

と対角化されます.

応用上は, 対称行列を扱うことがよくあります.

定義 2.8 (対称行列)

行列 A が**対称** (symmetric) であるとは,

$$A^\top = A$$

となることをいいます. □

対称行列の固有値については, 次の定理 2.5 に示す性質が知られています.

定理 2.5 (対称行列の固有値)

対称行列 A の固有値はすべて実数になります. □

また対称行列を対角化する際は, 固有ベクトルを並べてつくる行列 P を**直交行列** (orthogonal matrix), つまり $P^\top P = PP^\top = I_n$ となるようにすることができます.

以上のことを利用して, 次の**問題 6** を解いてみましょう.

問題 6 (対角化)

行列 A が

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 4 & -1 \\ 1 & -1 & 4 \end{pmatrix}$$

[*15] つまり, 固有ベクトルが $\mathbb{R}^n$ の基底となる, ということです.

となっています. 以下の問に答えてください.

(1) 行列 A のすべての固有値を求めてください.
(2) (1) で求めたそれぞれの固有値に対応する固有ベクトル $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ を互いに正規直交するように選んでください.
(3) 行列 $P = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3)$ が直交行列であることを確認してください.
(4) 行列 A を直交行列 P によって対角化してください.

ヒント

(1) 行列 A の固有方程式を解けば良いです.
(2) 実は, 異なる固有値に対応する固有ベクトルは互いに直交します. 一方で, 同じ固有値に対応する固有ベクトルを正規直交するように選ぶ必要がありますが, これは Gram–Schmidt の正規直交化法を利用すればできます.
(3) $P^\top P = PP^\top = I_3$ となるかどうかを確認するだけです.
(4) 問題の前に説明した方法で対角化をすれば良いです.

3 確率論と統計学の復習

機械学習では多数のデータを扱う必要があり, 統計学とは切っても切れない関係にあります. 統計学は確率論と密接に関係しており, ある程度は確率論の理解も必要になります. 以下では, 確率論と統計学の基礎的な部分を復習してみましょう.

3.1 確率変数と確率分布

確率的に値が変わる変数を**確率変数** (random variable) と呼びます[*16]. 確率変数 X が値 x を取る確率や, 確率変数 X が a から b までの値を取る確率をそれぞれ $P(X = x)$, $P(a \leq X \leq b)$ などと書きます. この P を確率変数 X の従う**確率分布** (probability distribution) といいます[*17]. 確率分布が決まると, 確率変数 X の分布関数が定義されます[*18].

定義 3.1 (分布関数)
確率変数 X の**分布関数** (distribution function) F_X を

$$F_X(x) := P(X \leq x)$$

[*16] 実は, これは正確な定義ではありません. きちんと定義しようとすると測度論と呼ばれる数学の分野の知識が必要となりますが, このテキストでは深入りしないことにします.
[*17] この定義も正確ではありません.
[*18] 逆に, 確率変数の分布関数が決まればその変数の従う確率分布が決定されます. つまり, 確率分布と分布関数は 1 対 1 に対応します.

で定義します. □

確率変数の取りうる値が離散値か連続値かで, 離散値確率変数, 連続値確率変数と区別します. 離散値確率変数の従う分布は, その変数の確率質量関数によって決定されます.

定義 3.2 (確率質量関数)
離散値確率変数 X が取りうる値は $x_1, x_2, \ldots$ $(x_1 < x_2 < \cdots)$ であるとします. このとき,

$$p_X(x_i) = P(X = x_i)$$

となるような関数 p_X を考えることができます. この関数 p_X を X の**確率質量関数** (probability mass function) といいます. □

連続値確率変数については, 確率密度関数が存在することがあります. やはり, 連続値確率変数の従う分布は, その変数の確率密度関数によって決定されます.

定義 3.3 (確率密度関数)
連続値確率変数 X について,

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u, \quad \left(\text{つまり, } f_X(x) = \frac{\mathrm{d}F_X}{\mathrm{d}x}(x)\right)$$

となる非負値関数 $f_X(x)$ が存在することがあります[*19]. この関数 $f_X(x)$ を X の**確率密度関数** (probability density function) といいます. □

よく誤解されやすいのですが, 確率密度関数の $x = a$ での値 $f_X(a)$ は $X = a$ となる確率 $P(X = a)$ とは異なります. 連続値確率変数 X が a から b の間の値を取る確率は,

$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F_X(b) - F_X(a) \\ &= \int_a^b f_X(x)\,\mathrm{d}x \end{aligned}$$

で求まります. このため, この式で $a = b$ とすると

$$P(X = a) = \int_a^a f_X(x)\,\mathrm{d}x = 0$$

となり, 連続値確率変数 X がピッタリ a の値を取る確率は必ず 0 になることがわかります.

それでは, 問題を解いてここまでの事項を確認してみましょう. まずは, 離散確率分布に関する**問題 7** を解いてみてください.

[*19] 存在しないこともありますが, 実用上はあまり深く考える必要はありません. 分布関数を微分したものが確率密度関数だと思っておけば十分でしょう.

問題 7 (二項分布, 幾何分布)
表が $\frac{1}{3}$ の確率で, 裏が $\frac{2}{3}$ の確率で出る歪んだコインがあります. 以下の問いに答えてください.

(1) コイン投げを n 回行って, 表が出る回数を X とします. 確率変数 X の確率質量関数 $p_X(x)$ を求めてください.
(2) コイン投げを表が初めて出るまで行って, 表が出るまでにコインを投げた回数を Y とします. 確率変数 Y の確率質量関数 $p_Y(y)$ を求めてください.

ヒント

(1) 各 $x = 0, \ldots, n$ に対し, x 回の表が出る確率 $p_X(x)$ を求めます. 反復試行の確率の計算を思い出しましょう. ちなみに, X の従う分布は**二項分布** (binomial distribution) と呼ばれる分布になります.
(2) 各 $y = 0, 1, \ldots$ に対し, y 回の表が出る確率 $p_Y(y)$ を求めます. これは, 連続で $y-1$ 回裏が出て, 最後に表が出る確率を考えれば良いです. ちなみに, Y の従う分布は**幾何分布** (geometric distribution) と呼ばれる分布になります.

続いて, 連続確率分布に関する**問題 8** を解いてみましょう.

問題 8 (指数分布)
分布関数が, $\lambda > 0$ を定数として

$$F(x) = \begin{cases} 1 - \mathrm{e}^{-\lambda x} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

で与えられる分布の確率密度関数 $f(x)$ を求めてください.

ヒント
確率密度関数は, 分布関数を微分することで得られます[*20]. ちなみに, この分布は**指数分布** (exponential distribution) と呼ばれる分布になります.

3.2 期待値と分散

確率変数 X のとる値について, 期待値と分散と呼ばれる量が定義されます.

[*20] なお, 原点 $x = 0$ では微分不可能ですが, 実は確率密度関数の 1 点での値はどのようにとっても問題ないので, 特に $f(0) = \lim_{x \to +0} F'(x)$ などとしておけば構いません.

定義 3.4 (期待値, 分散)

確率変数 X の**期待値** (expectation) あるいは**平均** (mean) を

$$\mathbb{E}[X] := \begin{cases} \sum_x x p_X(x) & (X \text{ が離散値確率変数}) \\ \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x & (X \text{ が連続値確率変数}) \end{cases}$$

で定義します ($\sum_x$ は取りうるすべての x についての和をとることを意味します). また, 確率変数 X の**分散** (variance) を, $\mu = \mathbb{E}[X]$ として

$$\mathbb{V}[X] := \mathbb{E}\left[(X-\mu)^2\right] = \begin{cases} \sum_x (x-\mu)^2 p_X(x) & (X \text{ が離散値確率変数}) \\ \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,\mathrm{d}x & (X \text{ が連続値確率変数}) \end{cases}$$

で定義します. □

期待値は X が平均的にどのような値を取るのかを表す指標で, 分散は X が平均値から平均的にどの程度ばらつくかを表す指標になっています.

期待値について, 次の公式はよく利用されます.

定理 3.1 (関数の期待値)

確率変数 X と関数[*21] h について,

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\,\mathrm{d}x$$

が成立します. □

また, 期待値の計算の際には期待値の**線形性** (linearity) と呼ばれる性質を利用すると便利な場合があります.

定理 3.2 (期待値の線形性)

確率変数 X と, 定数 a, b, c, 関数 f, g について,

$$\mathbb{E}[af(X) + bg(X) + c] = a\mathbb{E}[f(X)] + b\mathbb{E}[g(X)] + c$$

が成立します. □

さて, それでは期待値と分散に関する**問題 9** を解いてみましょう.

[*21] 正確には「Borel 可測関数」ですが, 実用上は特に気にする必要はありません. 以下では気にせず「関数」と書くことにします.

問題 9 (期待値と分散)

以下の問に答えてください.

(1) 一般の確率変数 X の分散について, $\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$ が成立することを証明してください.

(2) 確率変数 X が, 確率密度関数

$$f_X(x) = \begin{cases} 1 & (0 \leq x \leq 1) \\ 0 & (x < 0,\ 1 < x) \end{cases}$$

の分布に従っているとします. 確率変数 X の期待値と分散を求めてください.

(3) 確率変数 Y が, 確率密度関数

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)$$

の分布に従っているとします. 確率変数 Y の期待値と分散を求めてください.

ヒント

(1) 証明には, 期待値の線形性 (定理 3.2) を利用します.

(2) 期待値の計算は定義 3.4 の通りに行います. 分散の計算を定義どおり行っても良いですが, (1) で示したことを利用しても構いません. ちなみに, この分布は**一様分布** (uniform distribution) と呼ばれる分布です.

(3) (2) と同様に, 期待値と分散の計算は定義どおり行えば大丈夫です. 分散の計算では, $(\exp(-\frac{1}{2}y^2))' = -y\exp(-\frac{1}{2}y^2)$ であることに注意して部分積分をしたあと, **発展問題 1** の結果を使うとよいでしょう. ちなみに, この分布は**標準正規分布** (standard normal distribution) と呼ばれる分布で, 応用上とても重要な分布です.

3.3 多次元確率変数

同時に複数の確率変数を考える場合は, それらの同時分布を考える必要があります. 複数の確率変数のうちの一部の変数に着目したい場合は, 周辺分布関数を考えます.

定義 3.5 (同時分布関数と周辺分布関数)

二つの n 次元ベクトル $\boldsymbol{a}, \boldsymbol{b}$ について $\boldsymbol{a} \leq \boldsymbol{b}$ とは, 各成分が $a_i \leq b_i$ $(i = 1, \ldots, n)$ となることをいいます. 確率変数ベクトル $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ の**同時分布関数** (joint distribution function) $F_{\boldsymbol{X}}$ を

$$F_{\boldsymbol{X}}(\boldsymbol{x}) := P(\boldsymbol{X} \leq \boldsymbol{x})$$

で定義します. また, $\boldsymbol{X}_i = (X_1, \ldots, X_i)^\top$, $\boldsymbol{x}_i = (x_1, \ldots, x_i)^\top$ として,

$$F_{\boldsymbol{X}_i}(\boldsymbol{x}_i) := P(\boldsymbol{X}_i \leq \boldsymbol{x}_i)$$

を $\boldsymbol{X}_i$ の**周辺分布関数** (marginal distribution function) といいます[*22]. □

以下では, 連続値確率変数について扱うことにします.

1 次元の連続値確率変数の分布関数について確率密度関数が存在することがあったのと同様に, 連続値確率変数ベクトルの同時分布関数に対しても同時密度関数が存在することがあります.

定義 3.6 (同時確率密度関数と周辺確率密度関数)

連続値確率変数ベクトル $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ の同時分布関数が

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{\mathbb{R}^n} f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

となる n 変数非負値関数 $f_{\boldsymbol{X}}(\boldsymbol{x})$ が存在することがあります. この関数 $f_{\boldsymbol{X}}(\boldsymbol{x})$ を $\boldsymbol{X}$ の**同時確率密度関数** (joint probability density function) といいます. また, $\boldsymbol{X}_i = (X_1, \ldots, X_i)^\top$, $\boldsymbol{x}_i = (x_1, \ldots, x_i)^\top$, $\tilde{\boldsymbol{x}}_i = (x_{i+1}, \ldots, x_n)^\top$ として,

$$f_{\boldsymbol{X}_i}(\boldsymbol{x}_i) := \int_{\mathbb{R}^{n-i}} f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\tilde{\boldsymbol{x}}_i$$

を $\boldsymbol{X}_i$ の**周辺確率密度関数** (marginal probability density function) といいます. 周辺確率密度関数を求める操作を周辺化と呼びます. □

さて, 多次元の場合にも期待値を考えることができ, 次の定理が多用されます.

定理 3.3 (期待値)

確率変数ベクトル $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ について, その同時確率密度関数を $f_{\boldsymbol{X}}(\boldsymbol{x})$ とします. このとき, 実数値関数 $h(\boldsymbol{X})$ の期待値は

$$\mathbb{E}\left[h(\boldsymbol{X})\right] = \int_{\mathbb{R}^n} h(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

で計算できます. □

なお, 多次元の場合も期待値の線形性が成立します.

定理 3.4 (期待値の線形性)

確率変数ベクトル $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ と定数 a, b, c, 関数 f, g について,

$$\mathbb{E}\left[af(\boldsymbol{X}) + bg(\boldsymbol{X}) + c\right] = a\mathbb{E}\left[f(\boldsymbol{X})\right] + b\mathbb{E}\left[g(\boldsymbol{X})\right] + c$$

が成り立ちます. □

[*22] ここではじめ i 個の確率変数の周辺分布関数についてのみ書いていますが, 適当に選んだ i 個の確率変数の周辺分布を考えることができます.

多変数の場合の期待値をつかって, 次の共分散, 相関係数が定義できます. 相関係数は二つの確率変数にどの程度線形性があるかを表す指標です.

定義 3.7 (共分散と相関係数)

確率変数 X, Y の平均が μ_X, μ_Y, 分散が $\sigma_X^2, \sigma_Y^2 > 0$ であるとします. このとき, X と Y の**共分散** (covariance) を

$$\mathrm{Cov}\left[X, Y\right] := \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$$

で定義します. また, X と Y の**相関係数** (correlation coefficient) を

$$\rho\left[X, Y\right] := \frac{\mathrm{Cov}\left[X, Y\right]}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

で定義します. 確率変数 X, Y について, $\rho\left[X, Y\right] = 0$ となるとき, X と Y は**無相関** (uncorrelated) であるといいます. □

複数の確率変数を考える際, 独立性は特に重要です. 統計の多くの議論では, サンプルが独立で同分布に従っていることを仮定しています.

定義 3.8 (独立性)

確率変数ベクトル $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ の同時分布関数 $F_{\boldsymbol{X}}$ が, 各確率変数変数の周辺分布関数 $F_{X_1}, \ldots, F_{X_n}$ を用いて,

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

と表せるとき, 確率変数 $X_1, \ldots, X_n$ は**独立** (independent) であるといいます[*23]. 独立な確率変数 $X_1, \ldots, X_n$ がさらに同一の分布に従っているとき, すなわち,

$$F_{X_i}(x) = F_{X_j}(x) \quad (i \neq j)$$

となるとき, 確率変数 $X_1, \ldots, X_n$ は**独立同分布** (independent and identically distributed) であるといって, i.i.d. と表記します. □

独立であれば無相関になりますが, 逆は正しくないことに注意しましょう. つまり, 無相関でも独立でない例が存在します[*24]. 連続確率変数の独立性については, 以下の同値な条件も利用されます.

[*23] ここで述べている独立性は「確率変数の」独立性であることを注意しておきます. 「事象の」独立性というものもありますが, これは確率変数の独立性とは異なっています. 例えば, 52 枚あるトランプから 1 枚引いたカードのスートが「スペード」である事象を A, 「キング」のカードである事象を B とすると, 事象 A, B は独立です. ここで, 「スペード」を「ハート」, 「キング」を「クイーン」に置き換えたとしても事象の独立性は成り立ちます. つまり, スートの観測とカードの数の観測そのものが独立になっています. スートを X, カードの数を Y という確率変数でそれぞれ表せば, これは確率変数の独立性を意味します. この意味で, 確率変数の独立性は事象の独立性よりも広い概念になっています.

[*24] 例えば, X と $Y = X^2$ は, 線形関係が無いために無相関ですが, Y の作り方からも分かるように独立ではありません.

定理 3.5 (独立性の同値条件)

連続確率変数 $X_1, \ldots, X_n$ が独立であることと, 同時確率密度関数 $f_{\boldsymbol{X}}(\boldsymbol{x})$ が

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

と表せることは同値です. □

複数の確率変数を考える際, その一部の値が観測された状況で, 残りの確率変数の確率分布を考えることがあります.

定義 3.9 (条件付き確率密度関数)

確率変数ベクトル $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ のうち, 確率変数 $\boldsymbol{X}_i = (X_1, \ldots, X_i)^\top$ が $\boldsymbol{X}_i = \boldsymbol{x}_i = (x_1, \ldots, x_i)^\top$ と観測されたとします. 残りの $\tilde{\boldsymbol{X}}_i = (X_{i+1}, \ldots, X_n)^\top$ の確率密度関数を

$$f_{\tilde{\boldsymbol{X}}_i|\boldsymbol{X}_i}(\tilde{\boldsymbol{x}}_i|\boldsymbol{x}_i) := \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{\boldsymbol{X}_i}(\boldsymbol{x}_i)} \quad (f_{\boldsymbol{X}_i}(\boldsymbol{x}_i) > 0)$$

と定め, $\boldsymbol{X}_i = \boldsymbol{x}_i$ が与えられたときの $\tilde{\boldsymbol{X}}_i$ の**条件付き確率密度関数** (conditional probability density function) といいます. □

もし, $X_1, \ldots, X_n$ が独立であれば,

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}_i}(\boldsymbol{x}_i) f_{\tilde{\boldsymbol{X}}_i}(\tilde{\boldsymbol{x}}_i)$$

ですので,

$$f_{\tilde{\boldsymbol{X}}_i|\boldsymbol{X}_i}(\tilde{\boldsymbol{x}}_i|\boldsymbol{x}_i) = f_{\tilde{\boldsymbol{X}}_i}(\tilde{\boldsymbol{x}}_i)$$

となり, $\boldsymbol{x}_i$ の値に依りません. よって直観的には, 独立であるような確率変数は「関係がない」ということを表しています.

この節の内容を確認するために, **問題 10** を解いてみましょう.

問題 10 (独立性, 標本平均と標本分散)

確率変数 $X_1, \ldots, X_n$ は i.i.d. であるとします. 次の問に答えてください.

(1) 異なる確率変数 X_i, X_j $(i \neq j)$ の共分散について, $\mathrm{Cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$ であることを証明してください.

(2) 確率変数の積の期待値について, $\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n]$ であることを証明してください.

(3) 確率変数の和の分散について, $\mathbb{V}[X_1 + \cdots + X_n] = \mathbb{V}[X_1] + \cdots + \mathbb{V}[X_n]$ であることを証明してください.

(4) 確率変数 X_1 の平均と分散が $\mathbb{E}[X_1] = \mu$, $\mathbb{V}[X_1] = \sigma^2 > 0$ のとき, 確率変数 $X_1, \ldots, X_n$ の**標本平均** (sample mean)

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

の平均 $\mathbb{E}\left[\bar{X}\right]$ と分散 $\mathbb{V}\left[\bar{X}\right]$ を求めてください.

(5) 確率変数 X_1 の平均と分散が $\mathbb{E}[X_1] = \mu$, $\mathbb{V}[X_1] = \sigma^2 > 0$ のとき, 確率変数 $X_1, \ldots, X_n$ の**標本分散** (sample variance)

$$S^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

の平均 $\mathbb{E}\left[S^2\right]$ を求めてください.

ヒント

(1) 共分散の定義の式を展開して, 期待値の線形性を使います.

(2) 期待値の計算の式と, 確率密度関数を用いた独立性の式を利用します.

(3) まずは, 分散の定義式を使います. 分散の定義式から計算するには, $\mathbb{E}[X_1 + \cdots + X_n]$ を計算する必要があることに注意しましょう. 公式

$$(x_1 + \cdots + x_n)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i<j} x_i x_j$$

と, (2) の結果をうまく使いましょう.

(4) 平均の計算では期待値の線形性を用いるだけです. 分散の計算では, 期待値の線形性と (2) を利用します. (3) と同様の計算を行うと解けるかと思います.

(5) これも, 期待値の線形性と (2) を利用します. やや煩雑な計算が必要ですが, やはり (3) と計算は同様になります.

問題 10 では, 標本分散の分散の計算については問いませんでした. これは定義どおりやると大変煩雑な計算が必要になるからです. 次の**発展問題 4** で, もう少し簡単に計算をしてみましょう.

発展問題 4 (標本分散の分散)

確率変数 $X_1, \ldots, X_n$ は i.i.d. であり, 確率変数 X_1 の平均と分散が $\mathbb{E}[X_1] = \mu$, $\mathbb{V}[X_1] = \sigma^2 > 0$ とします. 確率変数 $X_1, \ldots, X_n$ の標本平均を

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

とし, 標本分散を

$$S^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

と書きます. 標本分散の分散を計算しましょう. 次の問に答えてください.

(1) **ガンマ関数** (gamma function) は,

$$\Gamma(x) := \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t$$

で定義されます. ガンマ関数の性質として, 0 と負整数以外の実数 x について $\Gamma(x+1) = x\Gamma(x)$ であることを証明してください.

(2) 自由度 k の**カイ二乗分布** (chi-squared distribution) は, 確率密度関数が

$$f(x) = \begin{cases} \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1}\mathrm{e}^{-\frac{x}{2}} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

で与えられます. 自由度 k のカイ二乗分布に従う確率変数 X の平均 $\mathbb{E}[X]$ と分散 $\mathbb{V}[X]$ を計算してください.

(3) 事実として, $\frac{nS^2}{\sigma^2}$ が自由度 $n-1$ のカイ二乗分布に従うことが知られています. これを用いて, 標本分散 S^2 の分散 $\mathbb{V}\left[S^2\right]$ を求めてください.

ヒント

(1) 部分積分をすることで証明することができます.

(2) 平均は定義どおり計算します. その際, (1) の結果を使います. 分散については, **問題 9** の (1) の結果を利用するのが良いでしょう.

(3) 事実を利用して式変形すれば結果を得ます.

3.4 最小二乗法

最後に, 最小二乗法について復習することにしましょう. ある変数 $x \in \mathbb{R}$ と別の変数 $y \in \mathbb{R}$ には $y = w_1 x + w_0$ という関係があることがわかっており, 係数 w_1, w_0 を統計的に決定するために, n 個のサンプルからなるデータセット

$$\mathcal{D} = \{\,(x_1, y_1), \ldots, (x_n, y_n)\,\}$$

を取得しました. このデータセット $\mathcal{D}$ を使って妥当な w_1, w_0 を決定しましょう[*25].

係数 w_1, w_0 の値が決まっているとき, サンプル $(x_i, y_i) \in \mathcal{D}$ に対する誤差は

$$|y_i - (w_1 x_i + w_0)|$$

[*25] 「妥当な」というのは, 「$\mathcal{D}$ 以外に取得したデータに対しても $y = w_1 x + w_0$ の関係式がある程度成り立つ」といった意味です. このことを「汎化誤差が小さい」といいます.

になっています. すると, データセット $\mathcal{D}$ 全体では誤差が

$$E_1(w_1, w_0) := \sum_{i=1}^{n} |y_i - (w_1 x_i + w_0)|$$

となります. すべてのサンプルに対しての誤差が小さいほど係数 w_1, w_0 は妥当だと考えられますので, w_1, w_0 を変数とみて最適化問題

$$\text{minimize} \quad E_1(w_1, w_0)$$

を解けば良いです. しかし, この問題の目的関数は凸関数ではあるものの, 微分ができないため解くのが難しくなっています.

そこで, (絶対値による) 誤差の和の代わりに, **二乗誤差** (squared error) の和

$$E_2(w_1, w_0) := \sum_{i=1}^{n} (y_i - (w_1 x_i + w_0))^2$$

を最小化することにしましょう. つまり,

$$\text{minimize} \quad E_2(w_1, w_0)$$

の最適化問題を考えます. このように, 二乗誤差を最小にするように係数 w_1, w_0 を決定する方法を**最小二乗法** (method of least square) といいます[*26].

さて, 実際に最小二乗法で係数を決定してみましょう.

問題 11 (単回帰)

ある変数 $x \in \mathbb{R}$ と別の変数 $y \in \mathbb{R}$ には $y = w_1 x + w_0$ という関係があることがわかっており, 係数 w_1, w_0 を統計的に決定するために, n 個のサンプルからなるデータセット

$$\mathcal{D} = \{ (x_1, y_1), \ldots, (x_n, y_n) \}$$

を取得しました. 最小二乗法により, 係数 w_1, w_0 を決定してください.

ヒント

上記の通り,

$$\text{minimize} \quad E_2(w_1, w_0)$$

の最適化問題を解けばよいです. この目的関数 $E_2(w_1, w_0)$ は w_1, w_0 について凸関数であり, w_1, w_0 について偏微分ができますので, 停留点を求めればそれが最小解になります (1.2 節を参照しましょう).

[*26] 一般には, $E_1(w_1, w_0)$ の最小化で得られる係数と, $E_2(w_1, w_0)$ の最小化で得られる係数は異なるはずです.

上では, x という 1 次元の変数と y という変数の関係を統計的に決定しました. これを単回帰分析と呼びます. 一方で, $\boldsymbol{x}$ という d 次元の変数と y という変数の関係を統計的に決定することもできます. これは重回帰分析と呼ばれます. 以下では, この重回帰分析について扱います.

ある変数 $\boldsymbol{x} \in \mathbb{R}^d$ と別の変数 $y \in \mathbb{R}$ には $y = \boldsymbol{w}^\top \boldsymbol{x} + w_0$ という関係があることがわかっており, 係数 $\boldsymbol{w}, w_0$ を統計的に決定するために, n 個のサンプルからなるデータセット

$$\mathcal{D} = \{ (\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n) \}$$

を取得しました. このデータセット $\mathcal{D}$ を使って妥当な $\boldsymbol{w}, w_0$ を決定しましょう. まず話を簡単にするために, 関係式 $y = \boldsymbol{w}^\top \boldsymbol{x} + w_0$ を以下のように書き直してみます:

$$y = \boldsymbol{w}^\top \boldsymbol{x} + w_0 = \begin{pmatrix} w_0 \\ \boldsymbol{w} \end{pmatrix}^\top \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}.$$

この変形から, 変数 $\boldsymbol{x}$ を

$$\tilde{\boldsymbol{x}} := \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

と変換し, $\boldsymbol{w}, w_0$ をまとめて

$$\tilde{\boldsymbol{w}} := \begin{pmatrix} w_0 \\ \boldsymbol{w} \end{pmatrix}$$

と書くことにすれば, 関係式は $y = \tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{x}}$ という定数項のない形で書けることがわかります. 記号が煩雑になるのを防ぐため, 以下では $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{x}}$ の代わりに $\boldsymbol{w}, \boldsymbol{x}$ と書いて, データセット $\mathcal{D}$ 内のデータに対しても上のような変換がなされているものとします[*27].

さて, 考える問題は関係式 $y = \boldsymbol{w}^\top \boldsymbol{x}$ の係数 $\boldsymbol{w}$ をデータセット $\mathcal{D}$ から統計的に決定することです. ここでも最小二乗法を使って決定することにしましょう. 最小化する目的関数は, 全データに対する二乗誤差の和

$$E_2(\boldsymbol{w}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i \right)^2$$

です. この関数も $\boldsymbol{w}$ について凸関数になっており, 各 w_i について偏微分が可能です. よって, 停留点を求めればそれが最小解になります.

ここで, 以下のような記法を導入することにしましょう.

定義 3.10 (勾配)

ベクトル変数 $\boldsymbol{x} \in \mathbb{R}^d$ とスカラー値関数 $f(\boldsymbol{x}) \in \mathbb{R}$ について, f の $\boldsymbol{x}$ における**勾配** (gradient) を

$$\frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) := \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\boldsymbol{x}) \end{pmatrix}$$

と定義します. □

[*27] つまり, 元のデータは $d-1$ 次元ベクトルで, この変換によって d 次元ベクトル $\boldsymbol{x}_i$ になっていると考えます.

この記法を用いると, 関数 f の停留点 $\bar{\boldsymbol{x}}$ は

$$\frac{\partial f}{\partial \boldsymbol{x}}(\bar{\boldsymbol{x}}) = \boldsymbol{0}$$

を満たす点になります. 勾配演算については, 線形性が成立します.

定理 3.6 (勾配の線形性)
ベクトル変数 $\boldsymbol{x} \in \mathbb{R}^d$ に関するスカラー値関数 $f(\boldsymbol{x}), g(\boldsymbol{x}) \in \mathbb{R}$ と定数 $a, b \in \mathbb{R}$ について, 線形性

$$\frac{\partial}{\partial \boldsymbol{x}}\left(af(\boldsymbol{x}) + bg(\boldsymbol{x})\right) = a\frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x}) + b\frac{\partial g}{\partial \boldsymbol{x}}(\boldsymbol{x})$$

が成り立ちます. □

また, 以下の公式が便利です.

定理 3.7 (勾配の公式)
ベクトル変数 $\boldsymbol{x} \in \mathbb{R}^d$, 定数ベクトル $\boldsymbol{c} \in \mathbb{R}^d$, $d \times d$ の対称行列 A に対して, 以下が成り立ちます:

(1) $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{c}^\top \boldsymbol{x}) = \boldsymbol{c}$.
(2) $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^\top A\boldsymbol{x}) = 2A\boldsymbol{x}$. □

更に, データセット内のデータを縦に並べた $n \times d$ の**計画行列** (design matrix)

$$X := \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}$$

とベクトル

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

を導入して, 目的関数 $E_2(\boldsymbol{w})$ を見通しの良い形に書き直します.

$$E_2(\boldsymbol{w}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{w}\right)^2 = \sum_{i=1}^{n} \left(y_i - (X\boldsymbol{w})_i\right)^2 = \|\boldsymbol{y} - X\boldsymbol{w}\|^2 .$$

結局, 解くべき問題は次の最適化問題です:

$$\text{minimize} \quad E_2(\boldsymbol{w}) = \|\boldsymbol{y} - X\boldsymbol{w}\|^2 .$$

それでは, 実際に係数の決定をしてみましょう.

発展問題 5 (重回帰)

ある変数 $\boldsymbol{x} \in \mathbb{R}^d$ と別の変数 $y \in \mathbb{R}$ には $y = \boldsymbol{w}^\top \boldsymbol{x}$ という関係があることがわかっており, 係数 $\boldsymbol{w}$ を統計的に決定するために, n 個のサンプルからなるデータセット

$$\mathcal{D} = \{\, (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \,\}$$

を取得しました. データを並べて得られる行列とベクトルを

$$X = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

とし, 係数 $\boldsymbol{w}$ を

$$E_2(\boldsymbol{w}) = \|\boldsymbol{y} - X\boldsymbol{w}\|^2$$

を最小化することで決定します (最小二乗法). 次の問に答えてください.

(1) 行列 $X^\top X$ が対称行列であることを確認してください.

(2) 目的関数 $E_2(\boldsymbol{w})$ を最小にする $\boldsymbol{w}^*$ が**正規方程式** (normal equation)

$$X^\top X \boldsymbol{w}^* = X^\top \boldsymbol{y}$$

を満たすことを示してください. さらに, 行列 $X^\top X$ が正則であるとき[*28], $\boldsymbol{w}^*$ はどのように表されるでしょうか.

(3) 目的関数 $E_2(\boldsymbol{w})$ の代わりに, $\lambda > 0$ を定数として

$$\tilde{E}_2(\boldsymbol{w}) = \|\boldsymbol{y} - X\boldsymbol{w}\|^2 + \lambda \|\boldsymbol{w}\|^2$$

を最小化することで $\boldsymbol{w}$ を推定することを考えます. このとき, 目的関数 $\tilde{E}_2(\boldsymbol{w})$ を最小にする $\hat{\boldsymbol{w}}$ はどのような関係式を満たすでしょうか. なお, 関数 $\tilde{E}_2(\boldsymbol{w})$ が凸関数であることを証明なしに利用しても構いません.

ヒント

(1) 対称行列の定義を示せば良いです. 転置の性質を使いましょう.

(2) 目的関数 $E_2(\boldsymbol{w})$ は凸関数ですので, 停留点を求めればそれが最小解です. 正規方程式は, 停留点の条件

$$\frac{\partial}{\partial \boldsymbol{w}} E_2(\boldsymbol{w}) \bigg|_{\boldsymbol{w}=\boldsymbol{w}^*} = \boldsymbol{0}$$

[*28] 行列 $X^\top X$ が正則であるためにはどのような条件が要るのか, 一度考えてみると良いですね.

から得られます. まずは

$$\|\boldsymbol{y} - X\boldsymbol{w}\|^2 = (\boldsymbol{y} - X\boldsymbol{w})^\top (\boldsymbol{y} - X\boldsymbol{w})$$

に注意して, これを展開して微分しましょう. 一般には X は正方行列でなく, "X^{-1}" のようなものは定義されていないことに注意しましょう (逆行列は正方行列に対してのみ定義されていました).

(3) やることは (2) と同じです. このように $\boldsymbol{w}$ を推定することを**リッジ回帰** (ridge regression) といいます.

参考文献

最後に, 今後勉強を進めていく上で役に立ちそうな本を紹介しておきます. ざっくりと機械学習で利用する数学について勉強したい人は, 次の本を図書館などで探して読んでみると良いでしょう:

[1] 石川 聡彦, 『人工知能プログラミングのための数学がわかる本』, KADOKAWA, 2018.

[2] 加藤 公一, 『機械学習のエッセンス 実装しながら学ぶ Python、数学、アルゴリズム』, SB クリエイティブ, 2018.

[1] は, 機械学習で利用される数学について基礎から説明してある本です. 平易に書かれているので, 数学が苦手な人でもとっつきやすいかと思います. [2] は, 基本的には Python というプログラミング言語による機械学習プログラミングに関する本ですが, 必要になるであろう数学について説明がなされています. この本のメインはどちらかといえば Python での実装なので, 研究室に入ってから Python でプログラミングをする場面でも役に立つと思います.

また, 微分積分学や線形代数学, 確率論・統計学に関する参考書を以下に示します. 微分積分の本はいろいろと出ていますが, 次の 3 冊を挙げておきます.

[3] 松野 陽一郎, 『なるほど! とわかる微分積分』, 東京図書, 2017.

[4] 齋藤 正彦, 『微分積分学』, 東京図書, 2006.

[5] 金谷 健一, 『これなら分かる最適化数学–基礎原理から計算手法まで』, 共立出版, 2005.

[3] は, 微分積分の重要な点が読みやすく, それでいて正確に書かれている点が良いです. [3] は, [3] よりは高尚な印象がありますが, 依然として読みやすく十分な詳しさがあります. [5] は, 正確には微分積分学の本ではなくて, 最適化数学に関する初学者向けの本です. 説明が丁寧で, 例も多く読みやすいです. 本テキストの最適化の節も, この本を参考に作りました.

線形代数の本もやはり多く世に出回っていますが, 次の 3 冊を挙げておきます.

[6] 松野 陽一郎, 『なるほど! とわかる線形代数』, 東京図書, 2017.

[7] 川久保 勝夫, 『線形代数学』(新装版), 日本評論社, 2010.

[8] 平岡 和幸, 堀 玄, 『プログラミングのための線形代数』, オーム社, 2004.

[6] は [3] と同じ著者による線形代数の本で, やはりこちらもわかりやすいです. [7] は [6] よりも難しく書いてある本ですが, 基本的なところはすべてカバーされていますので長く使えます. [8] は図が多く理解しやすい本です. 「プログラミングのための」とありますが, プログラミングがわからなくても問題なく読めます. 各概念の直観的な意味を掴むのに良いでしょう.

確率論・統計学についても近年は多く本が出ていますが, 初学者向けに 1 冊挙げるとすれば, [8] と同じ著者らによる [9] が良いと思います.

[9] 平岡 和幸, 堀 玄, 『プログラミングのための確率統計』, オーム社, 2009.

この本も, 図が多く平易に説明がなされているので, 初学者でも読みやすいと思います.

最後に, もう少し機械学習に関して知りたい人向けに, 以下の 2 冊を紹介しておきます.

[10] 大関真之, 『機械学習入門 ボルツマン機械学習から深層学習まで』, オーム社, 2016.
[11] 大関真之, 『ベイズ推定入門 モデル選択からベイズ的最適化まで』, オーム社, 2016.

[10], [11] ともに, 数式をほとんど使わずに深層学習や Bayes 推定について説明している本です. 気軽に楽しめる読み物として良いと思います.