数学の諸性質

中田和秀

東京工業大学 工学院 経営工学系

機械学習入門

http://www.iee.e.titech.ac.jp/~nakatalab/text/lecture.html

概要

ここでは機械学習入門の理解に必要な数学の記号や知識をまとめている。具体的には、行列や多変数（ベクトル）の話が中心である。なお、体系立てて説明しているわけではないため、1 から勉強したい人は専門書を読んでもらいたい。

目次：

記号の使い方：

- x, α など、実数やスカラー量はイタリック体を使う
- $\boldsymbol{x}, \boldsymbol{v}$ など、ベクトルはボールド体を使う
- $\boldsymbol{X}, \boldsymbol{A}$ など、行列は大文字のボールド体を使う
- $A := B$ は、B で A を定義する、B を A に代入することを意味する
- $[n]$ は n までのインデックスの集合を表し $[n] := \{1, 2, \cdots, n\}$

行列

記号の定義

- $\boldsymbol{A}^T \quad \cdots$ 転置
  行列 $\boldsymbol{A}$ に対し、ij 要素と ji 要素を入れ替えた行列
- $\boldsymbol{A}^{-1} \quad \cdots$ 逆行列 [a]
  行列 $\boldsymbol{A}$ に対し、$\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}$ を満たす $\boldsymbol{B}$

[a] 逆行列は存在するとは限らない。存在すれば一つといえる。

※ $\boldsymbol{I}$ は単位行列

用語の定義

- $\boldsymbol{A}$ が対称行列 $\overset{\text{def}}{\iff}$ $\boldsymbol{A} = \boldsymbol{A}^T$
- $\boldsymbol{A}$ が直交行列 $\overset{\text{def}}{\iff}$ $\boldsymbol{AA}^T = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$
- $\boldsymbol{A}$ が正則行列 $\overset{\text{def}}{\iff}$ $\boldsymbol{A}$ に逆行列が存在する

- 定義より、直交行列では $\boldsymbol{A}^{-1} = \boldsymbol{A}^T$ となることが分かる。

転置と逆行列に関する性質

- $(\boldsymbol{A}^T)^T = \boldsymbol{A}$
- $(\boldsymbol{A}\boldsymbol{B})^T = \boldsymbol{B}^T\boldsymbol{A}^T, \qquad (\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_m)^T = \boldsymbol{A}_m^T\boldsymbol{A}_{m-1}^T\cdots\boldsymbol{A}_1^T$
- $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$
- $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}, \qquad (\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_m)^{-1} = \boldsymbol{A}_m^{-1}\boldsymbol{A}_{m-1}^{-1}\cdots\boldsymbol{A}_1^{-1}$
- $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$ （この性質より $\boldsymbol{A}^{-T}$ と表記することがある）
- $\boldsymbol{A} = \boldsymbol{A}^T \implies \boldsymbol{A}^{-1} = (\boldsymbol{A}^{-1})^T$
- Sherman–Morrison の公式 $\boldsymbol{A} \in \mathbb{R}^{n\times n},\ \boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^n$

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{A}^{-1} - \frac{1}{1+\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}$$

- Sherman–Morrison–Woodbury の公式
（Woodbury の公式とも言う） $\boldsymbol{A} \in \mathbb{R}^{n\times n},\ \boldsymbol{U},\boldsymbol{V} \in \mathbb{R}^{n\times m},\ \boldsymbol{B} \in \mathbb{R}^{m\times m}$

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^T)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1} + \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}$$

固有値分解

行列 $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ に対する固有値 $\lambda \in \mathbb{R}$ と固有ベクトル $\boldsymbol{v} \in \mathbb{R}^n \quad (\boldsymbol{v} \neq \boldsymbol{0})$

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

線形独立な固有ベクトルが n 本ある場合、次のように行列 $\boldsymbol{\Lambda}, \boldsymbol{P}$ を定義

$$\boldsymbol{\Lambda} := \begin{pmatrix} \lambda_1, & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad \boldsymbol{P} := \left(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\right) \in \mathbb{R}^{n \times n}$$

固有値分解

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$$

$\boldsymbol{P} \in \mathbb{R}^{n \times n}$: 正則行列, $\quad \boldsymbol{\Lambda} \in \mathbb{R}^{n \times n}$ ：対角行列

※ 大半の行列で n 本の線形独立な固有ベクトルが取れるが、そうでない行列も少しあることに注意　→　次スライド

固有値分解の性質

固有値分解ができない（n 組の線形独立な固有ベクトルが存在しない）行列の例

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

その場合でも、似たような分解である Jordan（ジョルダン）分解はできるが、
用途が限られるので説明を省略

性質：$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$ としたとき、

- $\boldsymbol{A} + \alpha\boldsymbol{I} = \boldsymbol{P}(\boldsymbol{\Lambda} + \alpha\boldsymbol{I})\boldsymbol{P}^{-1}$

  行列 $\boldsymbol{A} + \alpha\boldsymbol{I}$ の固有値は $\boldsymbol{A}$ の固有値に α が足され、固有ベクトルは変わらない
- $\alpha\boldsymbol{A} = \boldsymbol{P}(\alpha\boldsymbol{\Lambda})\boldsymbol{P}^{-1}$

  行列 $\alpha\boldsymbol{A}$ の固有値は $\boldsymbol{A}$ の固有値が α 倍され、固有ベクトルは変わらない

正方行列に対する行列分解

機械学習入門では使わないが、有用な行列分解を紹介しておく

LU 分解と Schur 分解　　$\boldsymbol{A} \in \mathbb{R}^{n\times n}$

- LU 分解：

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U}$$

$\boldsymbol{P} \in \mathbb{R}^{n\times n}$：置換行列,　$\boldsymbol{L} \in \mathbb{R}^{n\times n}$：下三角行列,　$\boldsymbol{U} \in \mathbb{R}^{n\times n}$：上三角行列

- Schur（シュア）分解：

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{Q}^{-1} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{Q}^{T}$$

$\boldsymbol{Q} \in \mathbb{R}^{n\times n}$：直交行列,　$\boldsymbol{R} \in \mathbb{R}^{n\times n}$：上三角行列

※ 置換行列は各行各列に一つ 1 が入り、残りが 0 である行列。$\boldsymbol{P}\boldsymbol{A}$ は $\boldsymbol{A}$ の行が置換される。
※ 下三角行列は、対角部分とその左下部分にのみ非ゼロが入り、右上部分は 0 となる行列。

- LU 分解は、線形方程式を解くための Gauss の消去法（掃き出し法）と強い関係がある。多くの場合で $\boldsymbol{P}$ は必要ない。
- LU 分解と Schur 分解は全ての行列で可能。

一般行列に対する行列分解

機械学習入門では使わないが、有用な行列分解を紹介しておく

特異値分解と QR 分解　$\boldsymbol{A} \in \mathbb{R}^{m \times n}$　($m \geq n$ の場合を説明)

- 特異値分解：

$$\boldsymbol{A} = \boldsymbol{Q}_1 \boldsymbol{\Sigma} \boldsymbol{Q}_2$$

$\boldsymbol{Q}_1 \in \mathbb{R}^{m \times n}$：$\boldsymbol{Q}_1^T \boldsymbol{Q}_1 = \boldsymbol{I}$,　$\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$：非負の対角行列,　$\boldsymbol{Q}_2 \in \mathbb{R}^{n \times n}$：直交行列

- QR 分解：

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$$

$\boldsymbol{Q} \in \mathbb{R}^{m \times n}$：$\boldsymbol{Q}^T \boldsymbol{Q} = \boldsymbol{I}$,　$\boldsymbol{R} \in \mathbb{R}^{n \times n}$：上三角行列

※ 特異値分解の $\boldsymbol{Q}_1$ は正方行列ではなく、$\boldsymbol{Q}_1^T \boldsymbol{Q}_1 = \boldsymbol{I}$ であっても $\boldsymbol{Q}_1 \boldsymbol{Q}_1^T \neq \boldsymbol{I}$ であることに注意。QR 分解の $\boldsymbol{Q}$ も同様。

- 特異値分解は行列のサイズが異なる別の分解形式もある（本質的には同じ）
- QR 分解は Gram–Schmidt の直交化法と強い関係がある。
- 特異値分解と QR 分解は全ての行列で可能。

対称行列

対称行列の場合

- 固有値は必ず実数[1]
- 必ず固有値分解ができる
- 固有ベクトルは互いに直交するように取ることが可能

対称行列の固有値分解

$\boldsymbol{A}$：対称行列

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{T}$$

$\boldsymbol{Q} \in \mathbb{R}^{n\times n}$: 直交行列, $\quad \boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$：対角行列

- この分解は Schur 分解にもなっている

[1] 行列の各要素が実数であっても一般に固有値は複素数となるが、対称行列では実数に限定できる

半正定値

半正定値の定義

対称行列 $\boldsymbol{A}$ が半正定値であることの同値な条件：

- $\forall \boldsymbol{v} \in \mathbb{R}^n, \quad \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \geq 0$
- $\boldsymbol{A}$ の全ての固有値が非負
- $\exists \boldsymbol{B}, \ \boldsymbol{A} = \boldsymbol{B}^T \boldsymbol{B}$

同値な条件なので、どれか一つを定義とできる。

似たような概念として正定値というものがある。

対称行列 $\boldsymbol{A}$ が正定値であることの同値な条件：

- $\forall \boldsymbol{v}(\neq \boldsymbol{0}) \in \mathbb{R}^n, \quad \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} > 0$
- $\boldsymbol{A}$ の全ての固有値が正
- $\exists \boldsymbol{B}$（ただし、列フルランク）, $\boldsymbol{A} = \boldsymbol{B}^T \boldsymbol{B}$
- $\boldsymbol{A}$ が半正定値行列かつ正則行列

半正定値行列

半正定値行列の分解　$\boldsymbol{A} \in \mathbb{R}^{n \times n}$

- 固有値分解：

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{T}$$

$$\boldsymbol{Q} \in \mathbb{R}^{n \times n}\text{: 直交行列,} \quad \boldsymbol{\Lambda} \in \mathbb{R}^{n \times n}\text{：非負の対角行列}$$

- Cholesky（コレスキー）分解：

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{T}$$

$$\boldsymbol{L} \in \mathbb{R}^{n \times n}\text{：下三角行列}$$

- 平方根：次を満たすような半正定値行列 $\boldsymbol{A}^{1/2}$ が存在

$$\boldsymbol{A} = \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}$$

- 固有値分解は Schur 分解や特異値分解にもなっている
- Cholesky 分解は LU 分解の特殊ケース

半順序関係

対称行列 $\boldsymbol{X}, \boldsymbol{Y}$ に対して、$\boldsymbol{X} - \boldsymbol{Y}$ が半正定値行列のとき、

$$\boldsymbol{X} \succeq \boldsymbol{Y}$$

と表記する。$\succeq$ は半順序関係となる。

半順序関係

- $\boldsymbol{X} \succeq \boldsymbol{X}$ 反射律
- $\boldsymbol{X} \succeq \boldsymbol{Y},\ \boldsymbol{Y} \succeq \boldsymbol{X} \quad \Longrightarrow \quad \boldsymbol{X} = \boldsymbol{Y}$ 反対称律
- $\boldsymbol{X} \succeq \boldsymbol{Y},\ \boldsymbol{Y} \succeq \boldsymbol{Z} \quad \Longrightarrow \quad \boldsymbol{X} \succeq \boldsymbol{Z}$ 推移律

半正定値行列に対する相加相乗平均

$$\begin{aligned} \frac{1}{2}(\boldsymbol{A} + \boldsymbol{B}) &\succeq \boldsymbol{A}^{1/2}(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2})^{1/2}\boldsymbol{A}^{1/2} \\ &= \boldsymbol{B}^{1/2}(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2})^{1/2}\boldsymbol{B}^{1/2} \end{aligned}$$

多変数関数の微分

$f : \mathbb{R}^n \to \mathbb{R}$ に対する勾配

$$\nabla f(\boldsymbol{x}) := \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix}^T \in \mathbb{R}^n.$$

$f : \mathbb{R}^n \to \mathbb{R}$ に対するヘッセ行列（二階微分）

$$\nabla^2 f(\boldsymbol{x}) := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

部分的な微分

$f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ に対し、次のように部分的な勾配を定義する。

$$\nabla_x f(\boldsymbol{x}, \boldsymbol{y}) := \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_m} \end{pmatrix}^T \in \mathbb{R}^m,$$
$$\nabla_y f(\boldsymbol{x}, \boldsymbol{y}) := \begin{pmatrix} \frac{\partial f}{\partial y_1} & \frac{\partial f}{\partial y_2} & \cdots & \frac{\partial f}{\partial y_n} \end{pmatrix}^T \in \mathbb{R}^n.$$

ベクトル $\boldsymbol{a}$ と対称行列 $\boldsymbol{A}$ に対して、次の関係が成り立つ。

$$f(\boldsymbol{x}) := \boldsymbol{a}^T \boldsymbol{x} \quad \text{のとき}$$
$$\nabla f(\boldsymbol{x}) = \boldsymbol{a}, \quad \nabla^2 f(\boldsymbol{x}) = \boldsymbol{O}$$

$$f(\boldsymbol{x}) := \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \quad \text{のとき}$$
$$\nabla f(\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x}, \quad \nabla^2 f(\boldsymbol{x}) = 2\boldsymbol{A}$$

テーラー展開

$\tilde{x}$, $\tilde{\boldsymbol{x}}$ を中心としたテーラー展開

１変数の場合：

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(\tilde{x})(x - \tilde{x})^k$$
$$= f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}) + \frac{1}{2} f''(\tilde{x})(x - \tilde{x})^2 + \cdots$$

※ $f^{(k)}$ は f の k 階微分を意味する

多変数の場合：

$$f(\boldsymbol{x}) = f(\tilde{\boldsymbol{x}}) + \nabla f(\tilde{\boldsymbol{x}})^T (\boldsymbol{x} - \tilde{\boldsymbol{x}}) + \frac{1}{2} (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \nabla^2 f(\tilde{\boldsymbol{x}})(\boldsymbol{x} - \tilde{\boldsymbol{x}}) + \cdots$$

※ 上のように $f(x), f(\boldsymbol{x})$ が無限和で表されるには少し条件が必要

行列での微分

$f:\mathbb{R}^{m\times n}\to\mathbb{R}$ に対する勾配

$$\nabla f(\boldsymbol{X}) := \begin{pmatrix} \dfrac{\partial f}{\partial X_{11}} & \dfrac{\partial f}{\partial X_{12}} & \cdots & \dfrac{\partial f}{\partial X_{1n}} \\ \dfrac{\partial f}{\partial X_{21}} & \dfrac{\partial f}{\partial X_{22}} & \cdots & \dfrac{\partial f}{\partial X_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f}{\partial X_{m1}} & \dfrac{\partial f}{\partial X_{m2}} & \cdots & \dfrac{\partial f}{\partial X_{mn}} \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

性質 1：

$$\nabla \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{a}\boldsymbol{b}^T$$

$$\nabla \boldsymbol{a}^T\boldsymbol{X}^{-1}\boldsymbol{b} = -\boldsymbol{X}^{-T}\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{X}^{-T}$$

$$\nabla \log\det\boldsymbol{X} = \boldsymbol{X}^{-T}$$ (log 行列式の値は正とする)

ただし、$\boldsymbol{X}^{-T} := (\boldsymbol{X}^T)^{-1} = (\boldsymbol{X}^{-1})^T$

行列での微分の性質

性質 2：

$$
\begin{aligned}
&\nabla \mathrm{Tr}(\boldsymbol{AXB}) = \boldsymbol{A}^T\boldsymbol{B}^T \\
&\nabla \mathrm{Tr}(\boldsymbol{AX}^{-1}\boldsymbol{B}) = -\boldsymbol{X}^{-T}\boldsymbol{A}^T\boldsymbol{B}^T\boldsymbol{X}^{-T} \\
&\nabla \log\det \boldsymbol{AXB} = \boldsymbol{A}^T(\boldsymbol{AXB})^{-T}\boldsymbol{B}^T \\
&\nabla \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{b} = (\boldsymbol{b}\boldsymbol{a}^T + \boldsymbol{a}\boldsymbol{b}^T)\boldsymbol{X} \\
&\nabla \mathrm{Tr}(\boldsymbol{AXBX}^T\boldsymbol{C}) = \boldsymbol{A}^T\boldsymbol{C}^T\boldsymbol{XB}^T + \boldsymbol{CAXB} \\
&\nabla \boldsymbol{a}^T\boldsymbol{X}^2\boldsymbol{b} = \boldsymbol{a}\boldsymbol{b}^T\boldsymbol{X}^T + \boldsymbol{X}^T\boldsymbol{a}\boldsymbol{b}^T \\
&\nabla \mathrm{Tr}(\boldsymbol{AXBXC}) = \boldsymbol{A}^T\boldsymbol{C}^T\boldsymbol{X}^T\boldsymbol{B}^T + \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{C}^T \\
&\nabla \log\det(\boldsymbol{AXBX}^T\boldsymbol{A}^T) = 2\boldsymbol{A}^T(\boldsymbol{AXBX}^T\boldsymbol{A}^T)^{-1}\boldsymbol{AXB}
\end{aligned}
$$

※ $\log$ を取る行列の行列式の値は正で、最後の式で $\boldsymbol{B}$ は正定値とする。

ただし、$\boldsymbol{X}^{-T} := (\boldsymbol{X}^T)^{-1} = (\boldsymbol{X}^{-1})^T$

多変数ベクトル値関数の微分

出力（従属変数）がベクトルとなる関数：

$$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$$

偏微分（ヤコビ行列）　f_i は、$\boldsymbol{f}(\boldsymbol{x})$ の i 番目の要素を表す

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

- $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ とすると、$\dfrac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ とも書ける
- $m = 1$ の時、∇f は縦ベクトルで、$\dfrac{\partial f}{\partial \boldsymbol{x}}$ は横ベクトルなので、転置の関係となることに注意
- 勾配やヤコビ行列の表記法は色々あるが、本講義では ∇ と ∂ を使う

合成関数の微分

$$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m, \quad \boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^l$$

連鎖律

合成関数 $\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ の微分

$$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{f}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$$

$$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \in \mathbb{R}^{l \times n}, \quad \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{f}} \in \mathbb{R}^{l \times m}, \quad \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \in \mathbb{R}^{m \times n}$$

偏微分の例：

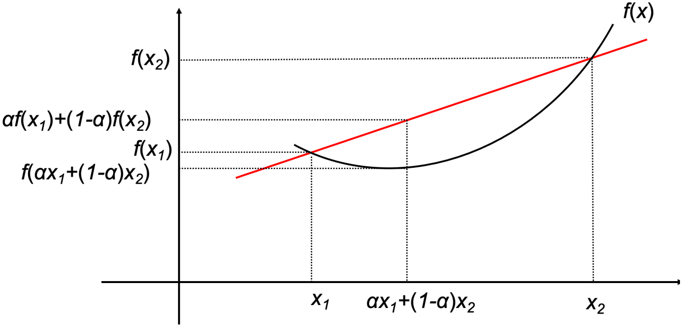
- $\boldsymbol{f}(\boldsymbol{x}) := \boldsymbol{A}\boldsymbol{x}$ のとき、$\dfrac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \boldsymbol{A}$
- $f(\boldsymbol{x}) := \dfrac{1}{1 + \exp\{-\boldsymbol{w}^T \boldsymbol{x}\}}$ のとき、
  $$\frac{\partial f}{\partial \boldsymbol{x}} = \frac{1}{1 + \exp\{-\boldsymbol{w}^T \boldsymbol{x}\}} \frac{1}{1 + \exp\{\boldsymbol{w}^T \boldsymbol{x}\}} \boldsymbol{w}^T$$

凸関数

凸関数

$f : \mathbb{R}^n \to \mathbb{R}$ が凸関数であるとは、次の条件を満たすことをいう。

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n, \quad \forall \alpha \in [0,1], \quad f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \le \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2).$$



- 2 階連続微分可能な $f(\boldsymbol{x})$ において、

  $f(\boldsymbol{x})$ が凸関数 $\iff$ $\forall \boldsymbol{x}, \ \nabla^2 f(\boldsymbol{x})$ が半正定値

確率変数

確率変数： $x, y \in \mathbb{R}$

- 平均： $\mathbb{E}[x] \in \mathbb{R}$
- 分散： $\mathbb{V}[x] := \mathbb{E}[(x - \mathbb{E}[x])^2] \in \mathbb{R}$
- 共分散： $\mathsf{Cov}[x, y] := \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \in \mathbb{R}$

性質

- $\mathbb{V}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$
- $\mathbb{E}[\alpha x + \beta] = \alpha \mathbb{E}[\boldsymbol{x}] + \beta$
  $\mathbb{V}[\alpha x + \beta] = \alpha^2 \mathbb{V}[\boldsymbol{x}]$
- $\mathbb{E}[\alpha x + \beta y] = \alpha \mathbb{E}[x] + \beta \mathbb{E}[y]$
  $\mathbb{V}[\alpha x + \beta y] = \alpha^2 \mathbb{V}[x] + 2\alpha\beta \mathsf{Cov}[x, y] + \beta^2 \mathbb{V}[y]$
- x と y が独立の場合、
  $\mathsf{Cov}[x, y] = 0$
  $\mathbb{V}[\alpha x + \beta y] = \alpha^2 \mathbb{V}[x] + \beta^2 \mathbb{V}[y]$

多次元の確率変数

多次元の確率変数：$\boldsymbol{x} := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$

- 平均ベクトル： $\mathbb{E}[\boldsymbol{x}] \in \mathbb{R}^n$
- 分散・共分散行列： $\mathbb{V}[\boldsymbol{x}] \in \mathbb{R}^{n \times n}$

$$\mathbb{E}[\boldsymbol{x}] := \begin{pmatrix} \mathbb{E}[x_1] \\ \mathbb{E}[x_2] \\ \vdots \\ \mathbb{E}[x_n] \end{pmatrix}, \quad \mathbb{V}[\boldsymbol{x}] := \begin{pmatrix} \mathbb{V}[x_1] & \mathrm{Cov}[x_1, x_2] & \cdots & \mathrm{Cov}[x_1, x_n] \\ \mathrm{Cov}[x_2, x_1] & \mathbb{V}[x_2] & \cdots & \mathrm{Cov}[x_2, x_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_n, x_1] & \mathrm{Cov}[x_n, x_2] & \cdots & \mathbb{V}[x_n] \end{pmatrix}$$

- 共分散行列： $\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] \in \mathbb{R}^{n \times n}$

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] := \begin{pmatrix} \mathrm{Cov}[x_1, y_1] & \mathrm{Cov}[x_1, y_2] & \cdots & \mathrm{Cov}[x_1, y_n] \\ \mathrm{Cov}[x_2, y_1] & \mathrm{Cov}[x_2, y_2] & \cdots & \mathrm{Cov}[x_2, y_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_n, y_1] & \mathrm{Cov}[x_n, y_2] & \cdots & \mathrm{Cov}[x_n, y_n] \end{pmatrix}$$

多次元確率変数の性質

確率変数 $\boldsymbol{x} \in \mathbb{R}^n$

- $$\begin{aligned}\mathbb{V}[\boldsymbol{x}] &= \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T] \\ &= \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T\end{aligned}$$
- 定数 $\boldsymbol{A} \in \mathbb{R}^{m \times n},\ \boldsymbol{b} \in \mathbb{R}^m$ に対し、
  $$\mathbb{E}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] + \boldsymbol{b} \in \mathbb{R}^m$$
  $$\mathbb{V}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{V}[\boldsymbol{x}]\boldsymbol{A}^T \in \mathbb{R}^{m \times m}$$
- $\boldsymbol{x}$ の各要素が独立の場合、$\mathbb{V}[\boldsymbol{x}]$ は対角行列

確率変数 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

- $\mathbb{E}[\alpha\boldsymbol{x} + \beta\boldsymbol{y}] = \alpha\mathbb{E}[\boldsymbol{x}] + \beta\mathbb{E}[\boldsymbol{y}]$
- $\mathbb{V}[\alpha\boldsymbol{x} + \beta\boldsymbol{y}] = \alpha^2\mathbb{V}[\boldsymbol{x}] + \alpha\beta\mathsf{Cov}[\boldsymbol{x}, \boldsymbol{y}] + \alpha\beta\mathsf{Cov}[\boldsymbol{y}, \boldsymbol{x}] + \beta^2\mathbb{V}[\boldsymbol{y}]$
- $\boldsymbol{x}$ と $\boldsymbol{y}$ が独立の場合、$\mathbb{V}[\alpha\boldsymbol{x} + \beta\boldsymbol{y}] = \alpha^2\mathbb{V}[\boldsymbol{x}] + \beta^2\mathbb{V}[\boldsymbol{y}]$

多変量正規分布

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$：平均ベクトル $\boldsymbol{\mu}$、分散共分散行列 $\boldsymbol{\Sigma}$ の正規分布

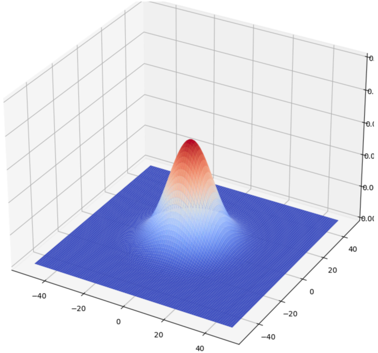
多変量正規分布

$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ のとき、確率密度関数は

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\det(\boldsymbol{\Sigma})^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right\}$$

特徴

- Gauss 分布とも呼ばれる
- 釣鐘型で対称性のある分布
- 多くの現象が当てはまる



多変量正規分布の性質

- 線形変換に対する再生性

  $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ のとき、$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \sim N(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$

- 和に対する再生性

  $\boldsymbol{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\boldsymbol{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ のとき、
  $\boldsymbol{x}_1 + \boldsymbol{x}_2 \sim N(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$

- 中心極限定理

  $\boldsymbol{x}_i$ が平均 $\boldsymbol{\mu}$、分散共分散行列 $\boldsymbol{\Sigma}$ となる同一の分布に従う時、

  $$\lim_{k\to\infty} \sum_{i=1}^{k} \frac{\boldsymbol{x}_i - \boldsymbol{\mu}}{\sqrt{k}} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

- １変数の正規分布との関係

  $x_i \sim N(0, \sigma^2)$ $(i \in [n])$ が独立のとき、$\boldsymbol{x} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ （$\boldsymbol{I}$ は単位行列）

ベイズ統計

事象 A_i は排反で、和が全事象となるとする

ベイズの定理

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_j P(A_j)P(B|A_j)}$$

$P(A_i)$：事象 A_i の事前確率
$P(A_i|B)$：事象 B の元での事象 A_i の事後確率

証明

$P(B)P(A_i|B) = P(A_i \cap B) = P(A_i)P(B|A_i)$ の両辺を $P(B)$ で割ると

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)}$$

$P(B) = \sum_j P(A_j \cap B) = \sum_j P(A_j)P(B|A_j)$

計算例

問題設定

- ある商品を購入する確率：20 代は 4%, 30 代は 2%,40 代は 1%
- 年代の分布：20 代は 30%, 30 代は 40%,40 代は 30%
- ある人が商品を購入していることがわかった時、その人が 20 代, 30 代, 40 代である確率は？

事前確率
20 代： 30%
30 代： 40%
40 代： 30%

事後確率
20 代： $0.3 \times 0.04/(0.3 \times 0.04 + 0.4 \times 0.02 + 0.3 \times 0.01) = 52\%$
30 代： $0.4 \times 0.02/(0.3 \times 0.04 + 0.4 \times 0.02 + 0.3 \times 0.01) = 35\%$
40 代： $0.3 \times 0.01/(0.3 \times 0.04 + 0.4 \times 0.02 + 0.3 \times 0.01) = 13\%$

Jensen の不等式

Jensen（イェンセン）の不等式

x：確率変数，　$f(x)$：凸関数

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

証明：

有限離散分布の場合で説明する。
確率変数を x_i $(i \in [I])$ とし、x_i が取る確率を p_i とする。　$\sum_{i\in[I]} p_i = 1,\ p_i \geq 0$

$$f(\mathbb{E}[x]) := f\left(\sum_{x_i \in X} p(x_i)x_i\right) = f\left(\sum_{i\in[I]} p_i x_i\right)$$

$$\mathbb{E}[f(x)] := \sum_{x_i \in X} p(x_i)f(x_i) = \sum_{i\in[I]} p_i f(x_i)$$

Jensen の不等式

$I = 1$ のときは、

$$f(\mathbb{E}[x]) = \mathbb{E}[f(x)] = f(x_1)$$

$I = 2$ のときは、

$$\begin{aligned} f(\mathbb{E}[x]) &= f(p_1 f(x_1) + p_2 f(x_2)) \\ &= f(p_1 f(x_1) + (1 - p_2) f(x_2)) \\ &\leq p_1 f(x_1) + (1 - p_1) f(x_2) && (\because f(x) \text{ は凸関数}) \\ &= p_1 f(x_1) + p_2 f(x_2) \\ &= \mathbb{E}[f(x)] \end{aligned}$$

Jensen の不等式

$I = k$ のとき成り立つとしたとき、$I = k + 1$ で成り立つことを示す。

$$
\begin{aligned}
f(\mathbb{E}[x]) &= f\left(\sum_{i\in[k]} p_i x_i \ + \ p_{k+1}x_{k+1}\right) \\
&= f\left((1-p_{k+1})\sum_{i\in[k]} \frac{p_i}{1-p_{k+1}} x_i \ + \ p_{k+1}x_{k+1}\right) \\
&\leq (1-p_{k+1}) f\left(\sum_{i\in[k]} \frac{p_i}{1-p_{k+1}} x_i\right) \ + \ f(p_{k+1}x_{k+1}) \quad (\because f(x) \text{ は凸関数}) \\
&\leq (1-p_{k+1})\sum_{i\in[k]} \frac{p_i}{1-p_{k+1}} f(x_i) \ + \ p_{k+1}f(x_{k+1}) \qquad (\because f(x) \text{ は凸関数}) \\
&= \sum_{i\in[k]} p_i f(x_i) \ + \ p_{k+1}f(x_{k+1}) \\
&= \mathbb{E}[f(x)]
\end{aligned}
$$

KL ダイバージェンス

２つの確率分布間の距離：

KL ダイバージェンス (Kullback–Leibler divergence)

$$\text{離散分布} \quad KL(P||Q) := \sum_{x_i \in X} p(x_i) \log \frac{p(x_i)}{q(x_i)}$$

$$\text{連続分布} \quad KL(P||Q) := \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

性質：

- $KL(P||Q) \geq 0$
- $KL(P||Q) = 0$ のとき $P = Q$
- 一般に $KL(P||Q) \neq KL(Q||P)$ であり、距離の公理を満たさないことに注意。

距離の公理

- $d(x, y) \geq 0, \quad d(x, y) = 0 \Longleftrightarrow x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$

証明

Jensen の不等式を使うと $KL(P||Q) \geq 0$ が示せる。

以下は、離散分布の場合で証明。

$$
\begin{aligned}
KL(P||Q) &= \sum_{x_i \in X} p(x_i) \log \frac{p(x_i)}{q(x_i)} = \sum_{x_i \in X} p(x_i) \left(- \log \frac{q(x_i)}{p(x_i)} \right) \\
&= \mathbb{E}\left[- \log \frac{q(x_i)}{p(x_i)} \right] \geq - \log \mathbb{E}\left[\frac{q(x_i)}{p(x_i)} \right] \qquad (\because \text{Jensen の不等式}) \\
&= - \log \sum_{x_i \in X} p(x_i) \frac{q(x_i)}{p(x_i)} = - \log \sum_{x_i \in X} q(x_i) \\
&= - \log 1 = 0
\end{aligned}
$$