Goodfellow：深層学習

Part1：応用数学と機械学習の基礎

第２章：線形代数

ノルムの整理

- L^p ノルム（$p < \infty$）の一般的な定義は以下の通り。

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p \right)^{1/p}$$

- 機械学習でよく使用されるのは、L^2, L^1, L^∞ ノルム。

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_i |x_i|^2}$$

ユークリッドノルム

$$\|\boldsymbol{x}\|_1 = \sum_i |x_i|$$

$$\|\boldsymbol{x}\|_\infty = \max_i |x_i|$$

最大値ノルム

- フロベニウスノルム：行列の各成分の２乗の総和の正の平方根

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} (A_{i,j})^2}$$

固有値分解

- **固有値と固有ベクトル**
  - 正方行列 A に対して、次の式を満たす非ゼロベクトル $\boldsymbol{v}$ を A の**固有ベクトル**、スカラー λ を**固有値**という。
  - $\boldsymbol{v}$ が A の固有ベクトルなら、その定数倍も同じ固有値をもつ固有ベクトルである。
    → 固有ベクトルを考える際はノルムが 1 の単位固有ベクトルを基本的には考える。

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

- **固有値分解**（Goodfellow本で考える固有値分解）
  - 実対称行列 A の固有値と固有ベクトルの組み合わせを $(\lambda_n, \boldsymbol{v}^{(n)})$ とおく。
  - 行列 V と Λ を $\boldsymbol{V} = \left[\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\right]$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ とおく。

$$\begin{aligned}\boldsymbol{A} &= \lambda_1 \boldsymbol{v}^{(1)}\left(\boldsymbol{v}^{(1)}\right)^\top + \cdots + \lambda_n \boldsymbol{v}^{(n)}\left(\boldsymbol{v}^{(n)}\right)^\top \\ &= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top\end{aligned}$$

特異値分解

- $m \times n$ の行列 A は $m \neq n$ の時には固有値分解ができない。
  しかし、特異値分解は任意の実行列に対して可能である。
- $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ に対して、$\boldsymbol{U} \in \mathbb{R}^{m \times m}, \boldsymbol{D} \in \mathbb{R}^{m \times n}, \boldsymbol{V} \in \mathbb{R}^{n \times n}$ を用いて、
  以下のような積で表すことを**特異値分解**という。

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}$$

- $\boldsymbol{D}$ の対角成分を行列 $\boldsymbol{A}$ の**特異値**といい、
  $\boldsymbol{U}, \boldsymbol{V}$ の各列をそれぞれ**左特異ベクトル**、**右特異ベクトル**という。
- 行列 $\boldsymbol{A}$ の左特異ベクトルは行列 $\boldsymbol{A}\boldsymbol{A}^{\top}$ の固有ベクトルであり、
  行列 $\boldsymbol{A}\boldsymbol{A}^{\top}$ の固有値の平方根は行列 $\boldsymbol{A}$ の特異値と一致する。

トレース演算子

- 行列 A の対角成分の総和として、$\mathrm{Tr}(\boldsymbol{A})$ が定義される。
  このトレース演算子について以下の関係式が成立する。

① 行列 A のフロベニウスノルム： $\|\boldsymbol{A}\|_F = \sqrt{\mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^\top)} = \sqrt{\mathrm{Tr}(\boldsymbol{A}^\top\boldsymbol{A})}$

② 行列 A の転置行列のトレース： $\mathrm{Tr}(\boldsymbol{A}^\top) = \mathrm{Tr}(\boldsymbol{A})$

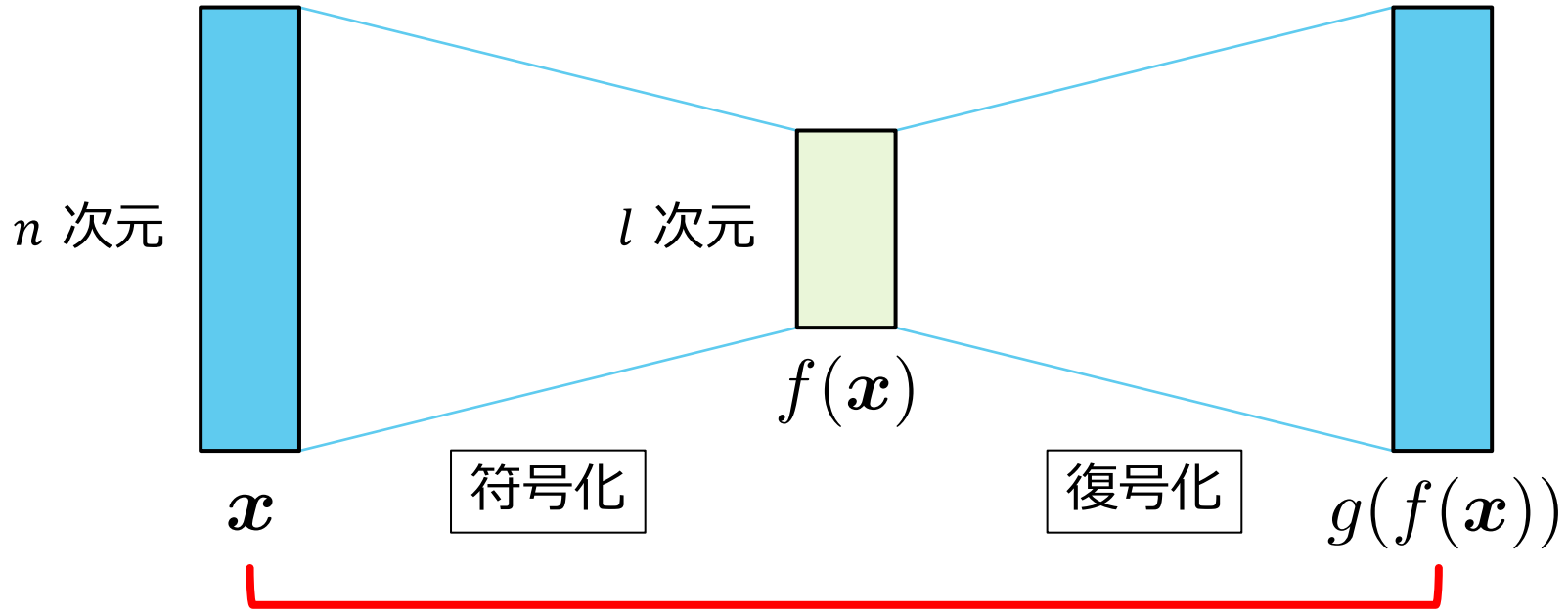
③ 3 つの正方行列の積のトレース： $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \mathrm{Tr}(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}) = \mathrm{Tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{A})$

④ 2 つの行列 A, B が正方行列でない場合の積のトレース：

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{Tr}(\boldsymbol{B}\boldsymbol{A})$$

主成分分析の導出 (1/11)

- $\mathbb{R}^n$ における m 個の点の集まり $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ の精度はそのままに低次元に圧縮することを考える。
- ここでは、上記の次元圧縮を符号化と復号化の枠組みで考える。



両者がおおよそ一致するような変換 f と g が得ることができれば、
適切な次元圧縮が可能になる。

主成分分析の導出 (2/11)

- **主成分分析 (PCA)**：復号化の変換として線形変換を使用。

$$g(\boldsymbol{c}) = \boldsymbol{D}\boldsymbol{c}$$

- 問題の簡単化のため、いくつかの制約を設ける。
  - Dの列が互いに直交することを仮定する。
  - Dとcを適当に定数倍したものも次元圧縮先としては適切な場合がありうるため、圧縮先の候補は無数に存在する。
    → Dの全ての列ベクトルが単位ベクトルであることを仮定する。

主成分分析の導出 (3/11)

- 主成分分析の定式化
  → 入力 x と復号化器の出力 g(c) の値が十分近くなるような
  x の符号化先 c = f(x) を探索すること

$$\boldsymbol{c}^* = \underset{\boldsymbol{c}}{\operatorname{argmax}} \|\boldsymbol{x} - g(\boldsymbol{c})\|_2^2$$

- $g(\boldsymbol{c}) = \boldsymbol{D}\boldsymbol{c}$ なので、上の最適化問題は以下のようになる。

$$\begin{aligned}
\boldsymbol{c}^* &= \underset{\boldsymbol{c}}{\operatorname{argmax}} \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{c}\|_2^2 \\
&= \underset{\boldsymbol{c}}{\operatorname{argmax}} \left(\boldsymbol{x}^\top \boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{c} + c^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{c}\right) \\
&= \underset{\boldsymbol{c}}{\operatorname{argmax}} \left(-2\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{c} + \boldsymbol{c}^\top \boldsymbol{c}\right)
\end{aligned}$$

cに依存しない

Dの各列が互いに直交し、
かつ単位ベクトルであるので、
$\boldsymbol{D}^\top \boldsymbol{D}$ は単位行列となる。

主成分分析の導出 (4/11)

- c^* を求める。
  - c に関する勾配が 0 になる点を求める。

$$\boldsymbol{c}^* = \underset{\boldsymbol{c}}{\operatorname{argmax}} \left(\boldsymbol{c}^\top \boldsymbol{c} - 2\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{c} \right)$$

$$\Rightarrow \quad \nabla_{\boldsymbol{c}} \left(\boldsymbol{c}^\top \boldsymbol{c} - 2\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{c} \right) = 2\boldsymbol{c} - 2\boldsymbol{D}^\top \boldsymbol{x} = \boldsymbol{0}$$

$$\Rightarrow \quad \boldsymbol{c}^* = \boldsymbol{D}^\top \boldsymbol{x}$$

【ベクトルの微分の公式】

$$\frac{\partial}{\partial \boldsymbol{x}} \left(\boldsymbol{a}^\top \boldsymbol{x} \right) = \boldsymbol{a}$$

$$\frac{\partial}{\partial \boldsymbol{x}} \left(\boldsymbol{x}^\top \boldsymbol{x} \right) = 2\boldsymbol{x}$$

- 以上より、復号化器が $g(\boldsymbol{c}) = \boldsymbol{D}\boldsymbol{c}$ の時に符号化器を $f(\boldsymbol{x}) = \boldsymbol{D}^\top \boldsymbol{x}$ と設計することが最適であることがわかった。
- この最適設計の下では $g\left(f(\boldsymbol{x})\right) = \boldsymbol{D}\boldsymbol{D}^\top \boldsymbol{x}$ であるため、行列 D は L2ノルム $(\|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{D}^\top \boldsymbol{x}\|_2)^2$ が最小となるように設計することが望ましい。

主成分分析の導出 (5/11)

- 行列 D の設計
  - m 個のデータの変換に対して共通の D を使用するため、
    全てのデータの処理に対して有効な行列を考える。
  - 「Dの列が互いに直交する」という制約条件を考慮する。

$$\boldsymbol{D}^* = \underset{\boldsymbol{D}}{\operatorname{argmin}} \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_j^{(i)} - \left(\boldsymbol{D}\boldsymbol{D}^\top \boldsymbol{x}^{(i)}\right)_j\right)^2}, \quad \text{subject to } \boldsymbol{D}^\top \boldsymbol{D} = \boldsymbol{I}$$

Dの列が互いに直交することに
対応する条件

入出力ベクトルの
各次元の一致

m個のデータ全ての総和をとることで、
全データに効果的なDを選択できるようにする。

主成分分析の導出 (6/11)

- 以下では $l = 1$ の場合を考えて、行列 D の設計方法を導出する。
  - 行列 D のサイズは $n \times l$ である。 $l = 1$ の場合、Dはn次元ベクトルとなる。
    （そこで D の代わりに d を用いて記すことにする。）
  - ベクトルのL2ノルムを使用すると以下の式と等価になる。
    → この式を変形することで主成分分析の導出法を示す。

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \sum_i \left(\| \boldsymbol{x}^{(i)} - \boldsymbol{d}\boldsymbol{d}^\top \boldsymbol{x}^{(i)} \|_2 \right)^2, \quad \text{subject to } \boldsymbol{d}^\top \boldsymbol{d} = 1$$

$\boldsymbol{d}^\top \boldsymbol{x}^{(i)}$ はスカラー：前に移動可能

スカラー：自身と転置したものが同一

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \sum_i \left(\| \boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)})^\top \boldsymbol{d}\boldsymbol{d} \|_2 \right)^2, \quad \text{subject to } \boldsymbol{d}^\top \boldsymbol{d} = 1$$

主成分分析の導出 (7/11)

- L2ノルムの部分を変形。

$$\left(\|\boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)})^\top \boldsymbol{d}\boldsymbol{d}\|_2\right)^2 = \sum_{j=1}^{n}\left[x_j^{(i)} - \left(\sum_{k=1}^{n} x_k^{(i)} d_k\right) d_j\right]^2 = \sum_{j=1}^{n}\left[x_j^{(i)} - \left(\sum_{k=1}^{n} x_k^{(i)} (d_k d_j)\right)\right]^2$$

  - 緑の下線部は以下の行列積の第 j 成分になる。

$$\begin{pmatrix} x_1^{(i)} & x_2^{(i)} & \cdots & x_n^{(i)} \end{pmatrix} \begin{pmatrix} d_1 d_1 & d_1 d_2 & \cdots & d_1 d_n \\ d_2 d_1 & d_2 d_2 & \cdots & d_2 d_n \\ \vdots & \vdots & \ddots & \vdots \\ d_n d_1 & d_n d_2 & \cdots & d_n d_n \end{pmatrix}$$

- X を $\boldsymbol{X}_{i,:} = \left(\boldsymbol{x}^{(i)}\right)^\top$ により定義すると、最適化問題は以下の式に帰着する。

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \left(\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top\|_F\right)^2, \quad \text{subject to } \boldsymbol{d}^\top \boldsymbol{d} = 1$$

主成分分析の導出 (8/11)

- フロベニウスノルムの部分を式変形する。

$$\underset{d}{\operatorname{argmin}}(\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top\|_F)^2$$

$$=\underset{\boldsymbol{d}}{\operatorname{argmin}}\,\mathrm{Tr}\left((\boldsymbol{X}-\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top)^\top(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top)\right)$$

$$=\underset{\boldsymbol{d}}{\operatorname{argmin}}\,\mathrm{Tr}\left((\boldsymbol{X}^\top-\boldsymbol{d}\boldsymbol{d}^\top\boldsymbol{X}^\top)(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top)\right)$$

$$=\underset{\boldsymbol{d}}{\operatorname{argmin}}\,\mathrm{Tr}\left(\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top-\boldsymbol{d}\boldsymbol{d}^\top\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{d}\boldsymbol{d}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top\right)$$

$$=\underset{d}{\operatorname{argmin}}\left(-2\mathrm{Tr}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top)+\mathrm{Tr}(\boldsymbol{d}\boldsymbol{d}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top)\right)$$

$$=\underset{d}{\operatorname{argmin}}\left(-2\mathrm{Tr}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top)+\mathrm{Tr}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top\boldsymbol{d}\boldsymbol{d}^\top)\right)$$

フロベニウスノルムは
トレースを使って書き換えられる

転置行列と元の行列の
トレースは同一

d に依存しない

中身の行列を巡回させても
トレースは不変

主成分分析の導出 (9/11)

- 前ページの式変形より、dに関する最適化問題は以下の式で表せる。

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \left(-2\mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top + \mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top \boldsymbol{d}\boldsymbol{d}^\top))\right), \quad \text{subject to } \boldsymbol{d}^\top \boldsymbol{d} = 1$$

- 制約条件とトレースの性質から次の式と等価である。

$$\begin{aligned}
\boldsymbol{d}^* &= \underset{\boldsymbol{d}}{\operatorname{argmin}} \left(-\mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top)\right), \quad \text{subject to } \boldsymbol{d}^\top \boldsymbol{d} = 1 \\
&= \underset{\boldsymbol{d}}{\operatorname{argmax}} \left(\mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^\top)\right), \quad \text{subject to } \boldsymbol{d}^\top \boldsymbol{d} = 1 \\
&= \underset{\boldsymbol{d}}{\operatorname{argmax}} \left(\mathrm{Tr}(\boldsymbol{d}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{d})\right), \quad \text{subject to } \boldsymbol{d}^\top \boldsymbol{d} = 1
\end{aligned}$$

主成分分析の導出 (10/11)

- Lagrangeの未定乗数法を利用する。

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmax}} \left(\operatorname{Tr}(\boldsymbol{d}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{d}) \right), \quad \text{subject to } \boldsymbol{d}^\top \boldsymbol{d} = 1$$

Lagrange関数

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmax}} \left(\operatorname{Tr}(\boldsymbol{d}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{d}) - \lambda(\boldsymbol{d}^\top \boldsymbol{d} - 1) \right)$$

- ベクトルの微分の公式を利用する。
  - 停留点 d^* は次の式を満たす。

$$2\left(\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{d}^* - \lambda \boldsymbol{d}^*\right) = 0 \implies \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{d}^* = \lambda \boldsymbol{d}^*$$

【ベクトルの微分の公式】

$$\frac{\partial}{\partial \boldsymbol{A}} \left(\operatorname{Tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^\top) \right) = \boldsymbol{A}(\boldsymbol{B} + \boldsymbol{B}^\top)$$

$$\frac{\partial}{\partial \boldsymbol{x}} \left(\boldsymbol{x}^\top \boldsymbol{x} \right) = 2\boldsymbol{x}$$

$$\frac{\partial}{\partial \boldsymbol{A}^\top} f(\boldsymbol{A}) = \left(\frac{\partial}{\partial \boldsymbol{A}} f(\boldsymbol{A}) \right)^\top$$

主成分分析の導出 (11/11)

- 以上より、最適なdを求める問題は $\boldsymbol{X}^\top \boldsymbol{X}$ の固有ベクトルに対応する。

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmax}} \left(\operatorname{Tr}(\boldsymbol{d}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{d}) - \lambda(\boldsymbol{d}^\top \boldsymbol{d} - 1) \right) \qquad \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{d}^* = \lambda \boldsymbol{d}^*$$

- $\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{d}^* = \lambda \boldsymbol{d}^*$ の条件をLagrange関数に代入すると、その値は λ となる。
  → Lagrange関数が最大となるのは λ が最大の時。
- つまり、**最適なdを求める問題は $X^\top X$ の最大固有値に対応する固有ベクトルを求める固有値問題と等価**である。