線形代数学 I: 第6回講義

データサイエンスに必要なベクトルと行列

中村 知繁

T. Nakamura | Juntendo Univ.

2025/02/08

はじめに

今回のテーマ:「1次元データとベクトルの和と積」

目標:

- 身近な1次元の数値データを「ベクトル」として捉える。
- ベクトル表現を用いて、データの基本的性質（平均、偏差、分散、標準偏差）を計算し、理解する。

はじめに

線形代数学の役割:

- ベクトルや行列は、抽象的な数学の道具であると同時に、データサイエンスや機械学習でデータを効率的に扱い、分析するための強力な言語。
- 今回の内容は、その第一歩。

本講義の学習目標:

1. 1次元データをベクトルとして表現する**意義**を理解し、実際に表現できる。
2. ベクトルの演算を用いてデータの**平均値**を計算できる。
3. ベクトルの演算を用いてデータの**偏差**を計算できる。
4. ベクトルの演算を用いてデータの**分散**を計算できる。

なぜデータをベクトルで表現するのか？

私たちの周りには、様々な数値データがあふれています。
（例：テストの点数、毎日の最高気温、血圧の測定値など）
これらは「1次元データ」の例です。

データ数が少ないうちは個別処理も可能ですが、データ数が増大すると（数百、数千...）、効率的かつ統一的な処理方法が必要になります。

→ **ここで「ベクトル」が役立ちます！**

なぜデータをベクトルで表現するのか？

ベクトル表現の主なメリット:

- **記述の簡潔さ**:
  多数の数値データ（例：100個）を、$\mathbf{x}$ という一つの記号で代表可能。
  数式やアルゴリズムの記述がスッキリし、見通しが良くなります。
- **計算の効率化**:
  ベクトル・行列演算に特化した計算ライブラリ（例：PythonのNumPy）の利用。
  大量のデータ処理を高速に実行できます。

なぜデータをベクトルで表現するのか？ (3/3)

ベクトル表現の主なメリット (続き):

- **幾何学的解釈の導入**:
  n個のデータをn次元空間内の「点」または「ベクトル」として解釈。
  データの分布や関係性について直感的な洞察を得やすくなります (特に多次元データで有効)。
- **多次元データへの自然な拡張性**:
  本講義は1次元データから開始。
  実世界のデータは多次元（複数の特徴量を持つ）が一般的。
  ベクトル・行列による表現は、多次元データ分析の基礎となります。

データベクトルの定義

1次元データをベクトルとしてどのように表現するかを定義しましょう。

“ **定義：データベクトル**
n個の観測値 $x_1, x_2, \ldots, x_n$ からなる1次元データセットは、n次元の列ベクトル $\mathbf{x}$ として以下のように表現できます。

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

このベクトル $\mathbf{x}$ を「データベクトル」と呼びます。 ”

- 本講義では、特に断りのない限り、データベクトルは**列ベクトル**とします。

例：体温データ

具体的な例を見てみましょう。
5人の学生A, B, C, D, Eの体温:
36.5°C, 36.8°C, 37.2°C, 36.4°C, 36.9°C

この1次元データセットは、5次元のデータベクトル $\mathbf{x}$ として次のように表現できます。

$$\mathbf{x} = \begin{pmatrix} 36.5 \\ 36.8 \\ 37.2 \\ 36.4 \\ 36.9 \end{pmatrix}$$

複数の数値データを一つのベクトルにまとめることで、計算や議論が扱いやすくなります。

全ての要素が1のベクトル (1 ベクトル) の定義と役割

データの統計量を計算する際に重要な役割を果たす特別なベクトルがあります。
それは、全ての要素が1であるベクトルで、$\mathbf{1}$（イチベクトル）と表記されます。

> **定義：全ての要素が1のベクトル (1 ベクトル)**
> n次元の「1ベクトル」$\mathbf{1}_n$ は、全ての要素が1である n次元の列ベクトルとして、以下のように定義されます。
>
> $$\mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$
>
> 次元 n が明らかな場合は、添え字 n を省略して $\mathbf{1}$ と書くこともあります。

例：3次元の1ベクトル $\mathbf{1}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

この $\mathbf{1}$ ベクトルは、データの総和や平均値をベクトル表記で簡潔に表す際に非常に便利です。

2. ベクトルを用いた記述統計量の計算

ここからは、データベクトルを用いて、基本的な記述統計量（平均値、偏差、分散、標準偏差、偏差値）を計算する方法を学びます。

特に**内積**がどのように活用されるかに注目してください。

2.1 平均値 : 定義

データの平均値（算術平均）は、データセット全体の代表的な値を示す最も基本的な統計量の一つです。

スカラ表記による定義

n個のデータ $x_1, x_2, \ldots, x_n$ の平均値 $\bar{x}$ は：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

「全てのデータの値を合計し、データの個数で割る」操作です。

2.1 平均値 : 定義

データベクトル $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ と n次元の1ベクトル $\mathbf{1}_n$ を考えます。

$\mathbf{1}_n$ の転置ベクトル $\mathbf{1}_n^T = \begin{pmatrix} 1 & \dots & 1 \end{pmatrix}$ と $\mathbf{x}$ の内積は、

$$\mathbf{1}_n^T \mathbf{x} = \sum_{i=1}^{n} x_i \quad (\text{データの総和})$$

となります。
したがって、平均値 $\bar{x}$ は、

“ **定義：データの平均値 (ベクトル表記)**

$$\bar{x} = \frac{1}{n}(\mathbf{1}_n^T \mathbf{x})$$

（内積記号 $\cdot$ を用いて $\bar{x} = \frac{1}{n}(\mathbf{1}_n \cdot \mathbf{x})$ とも書けます。） ”

ベクトルを用いると $\sum$ を使わずに平均値を表現でき、**数式がスッキリ**します。

2.1 平均値 : 例

5人の学生の体温データ: $\mathbf{x} = \begin{pmatrix} 36.5 \\ 36.8 \\ 37.2 \\ 36.4 \\ 36.9 \end{pmatrix}$

$n = 5$ なので、$\mathbf{1}_5^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}$ を用います。

$$\bar{x} = \frac{1}{5}(\mathbf{1}_5^T \mathbf{x}) = \frac{1}{5} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 36.5 \\ 36.8 \\ 37.2 \\ 36.4 \\ 36.9 \end{pmatrix}$$

$$= \frac{1}{5}(36.5 + 36.8 + 37.2 + 36.4 + 36.9)$$

$$= \frac{1}{5}(183.8) = 36.76$$

平均体温は 36.76℃ です。

2.2 偏差 : 定義

平均値はデータセットの中心的な傾向を示しますが、個々のデータが平均値からどれだけ離れているか（「偏差」）も重要です。

スカラ表記・ベクトル表記による定義

スカラ表記: 各データ点 x_i の平均値 $\bar{x}$ からの偏差 d_i は：

$$d_i = x_i - \bar{x}$$

ベクトル表記: 全ての偏差 d_i を要素とする「偏差ベクトル」$\mathbf{d}$ は：

$$\mathbf{d} = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix} = \mathbf{x} - \bar{x}\mathbf{1}_n = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}$$

$\bar{x}\ \mathbf{1}_n$ は、全要素が平均値 $\bar{x}$ のベクトル。

2.2 偏差 : 偏差の総和はゼロ

重要な性質: 「偏差の総和は常に0になる」 i.e., $\sum_{i=1}^{n} d_i = 0$

2.2 偏差：例

体温データ $\mathbf{x} = \begin{pmatrix} 36.5 \\ 36.8 \\ 37.2 \\ 36.4 \\ 36.9 \end{pmatrix}$, 平均値 $\bar{x} = 36.76$℃

$$\mathbf{d} = \mathbf{x} - \bar{x}\mathbf{1}_5 = \begin{pmatrix} 36.5 \\ 36.8 \\ 37.2 \\ 36.4 \\ 36.9 \end{pmatrix} - 36.76 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.26 \\ 0.04 \\ 0.44 \\ -0.36 \\ 0.14 \end{pmatrix}$$

偏差ベクトルの要素の和:
$(-0.26) + 0.04 + 0.44 + (-0.36) + 0.14 = 0$
確かに偏差の総和は0になります。

2.3 分散

- データがどの程度ばらついているか（散らばり具合）を一つの数値で表す代表的な指標が「分散」
- 分散は、各データ点の偏差を二乗し、それらを平均したもの。
- 偏差を二乗するのは、正負の値をそのまま合計すると0になるため、平均からの距離の大きさを正の値として評価するためです。

2.3 分散: 定義

スカラ表記: 分散 σ^2 (シグマ二乗) は：

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

2.3 分散: 定義

ベクトル表記: 偏差ベクトル $\mathbf{d}$ の自分自身との内積 ($\|\mathbf{d}\|^2$ に等しい) を用います。

$$\mathbf{d}^T\mathbf{d} = d_1^2 + \ldots + d_n^2 = \sum d_i^2$$

なので、

定義：データの分散 (ベクトル表記)

$$\sigma^2 = \frac{1}{n}(\mathbf{d}^T\mathbf{d}) = \frac{1}{n}\|\mathbf{d}\|^2$$

または、$\mathbf{d} = \mathbf{x} - \bar{x}\mathbf{1}_n$ を代入して、

$$\sigma^2 = \frac{1}{n}(\mathbf{x} - \bar{x}\mathbf{1}_n)^T(\mathbf{x} - \bar{x}\mathbf{1}_n)$$

- 分散大 → ばらつき大
- 分散小 → ばらつき小。

2.3 分散: 例

体温データの偏差ベクトル: $\mathbf{d} = \begin{pmatrix} -0.26 \\ 0.04 \\ 0.44 \\ -0.36 \\ 0.14 \end{pmatrix}$

$$\mathbf{d}^T\mathbf{d} = (-0.26)^2 + (0.04)^2 + (0.44)^2 + (-0.36)^2 + (0.14)^2$$

$$= 0.0676 + 0.0016 + 0.1936 + 0.1296 + 0.0196 = 0.412$$

分散 σ^2:

$$\sigma^2 = \frac{1}{5}(0.412) = 0.0824$$

この体温データの分散は 0.0824 です。

2.4 標準偏差

分散の単位は元のデータの単位の二乗となり直感的に解釈しづらいため、分散の正の平方根をとった「標準偏差」がよく用いられます。標準偏差の単位は、元のデータと同じになります。

2.4 標準偏差： 定義

定義：データの標準偏差

標準偏差 σ (シグマ) は、分散 σ^2 の正の平方根：

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

ベクトル表記では：

$$\sigma = \sqrt{\frac{1}{n}(\mathbf{d}^T\mathbf{d})} = \frac{1}{\sqrt{n}}\|\mathbf{d}\|$$

- 標準偏差大 → ばらつき大。
- 標準偏差小 → ばらつき小。

2.4 標準偏差:　例

体温データの分散 $\sigma^2 = 0.0824$ (℃²) でした。

標準偏差 σ は：

$$\sigma = \sqrt{0.0824} \approx 0.287054$$

有効数字を考慮し、標準偏差は約 0.287℃ となります。

平均体温 36.76℃、標準偏差 約 0.287℃ から、データは平均値から ±0.287℃ 程度の範囲に散らばっている、という目安が得られます。

2.5 分散の別公式

分散には、定義式以外にもう一つ便利な計算公式があります。

2.5 分散の別公式

定理：分散の計算公式 (別バージョン)

$$\sigma^2 = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \bar{x}^2$$

ベクトル表記：

$$\sigma^2 = \left(\frac{1}{n} \mathbf{x}^T \mathbf{x} \right) - \bar{x}^2$$

2.5 分散の別公式

分散の定義式から出発：

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

2.5 分散の別公式

2.5 分散の別公式: 例

体温データ $\mathbf{x} = \begin{pmatrix} 36.5 \\ \vdots \\ 36.9 \end{pmatrix}$, 平均値 $\bar{x} = 36.76$

$\mathbf{x}^T\mathbf{x} = \sum x_i^2$ を計算：

$36.5^2 = 1332.25$

$36.8^2 = 1354.24$

$37.2^2 = 1383.84$

$36.4^2 = 1324.96$

$36.9^2 = 1361.61$

$\mathbf{x}^T\mathbf{x} = 1332.25 + \ldots + 1361.61 = 6756.9$

2.5 分散の別公式: 例

「データの二乗の平均」:

$$\frac{1}{n}\mathbf{x}^T\mathbf{x} = \frac{1}{5}(6756.9) = 1351.38$$

「平均の二乗」:

$$\bar{x}^2 = (36.76)^2 = 1351.298$$

分散 σ^2:

$$\sigma^2 = 1351.38 - 1351.298 = 0.0824$$

2.6 偏差値

偏差値とは?

学力テストなどで使われ、集団の中で個人の成績がどの程度の位置にあるかを相対的に示す指標。

導入：なぜ偏差値が必要か？

異なるテスト間の単純な点数比較は難しい。

例: 数学80点 (平均85点) vs 英語70点 (平均60点)

テストの難易度 (平均点) や点数のばらつき (標準偏差) が異なると、相対的な位置を把握しにくい。

偏差値の役割:

平均が50、標準偏差が10となる共通の尺度に変換することで、比較しやすくする。

2.6 偏差値 : 定義と計算式

あるデータ x_i の偏差値 T_i は、平均 $\bar{x}$、標準偏差 σ のとき：

$$T_i = 10 \times \frac{x_i - \bar{x}}{\sigma} + 50$$

- x_i: 個人の得点など
- $\bar{x}$: データセット全体の平均値
- σ: データセット全体の標準偏差
- $\frac{x_i - \bar{x}}{\sigma}$: zスコア (標準得点)。平均から標準偏差の何倍分離れているかを示す。
  - $x_i = \bar{x} \implies z = 0$
  - $x_i = \bar{x} + \sigma \implies z = 1$
  - $x_i = \bar{x} - \sigma \implies z = -1$

2.6 偏差値 : 定義と計算式

- $10 \times z + 50$: zスコアを10倍して50を加え、平均50、標準偏差10の尺度に変換。
  - $z = 0$ (平均点) $\Longrightarrow$ $T = 10 \times 0 + 50 = 50$
  - $z = 1$ (平均点 + 1σ) $\Longrightarrow$ $T = 10 \times 1 + 50 = 60$
  - $z = -1$ (平均点 - 1σ) $\Longrightarrow$ $T = 10 \times (-1) + 50 = 40$

2.6 偏差値 :あるテストの得点から偏差値を計算する

5人の数学テスト結果 (100点満点):
A: 55点, B: 80点, C: 65点, D: 40点, E: 70点

計算結果 (再掲):
平均点: 62.00 点, 標準偏差: 14.00 点

- 学生A (55点): 偏差値 45.00
- 学生B (80点): 偏差値 62.86
- 学生C (65点): 偏差値 52.14
- 学生D (40点): 偏差値 34.29
- 学生E (70点): 偏差値 55.71

解釈: 平均(62点)なら偏差値50。平均+1σ(76点)なら偏差値60。

2.6 偏差値 : 性質と利点・注意点

性質と利点:

- **共通尺度**: 平均50、標準偏差10の尺度に変換される。
- **相対的位置の把握**: 集団内での自分の位置を客観的に評価。
- **異なるテスト間の比較**: 平均点やばらつきが異なるテスト結果を同じ土俵で比較可能。

注意点:

- **元のデータの分布形状**: 正規分布に近い場合に解釈しやすい。極端な偏りや外れ値が多い場合は注意。
- **集団のレベル**: あくまで集団内での相対位置。異なる集団間での絶対的な学力比較は困難。
- **1点の重み**: 元のテストの標準偏差により、偏差値1を上げるのに必要な素点が変わる。

3. 演習問題

ここまでの内容の理解を深めるために、いくつかの演習問題に取り組みましょう。
これらの問題は、手計算でも比較的容易に扱えるように数値を調整してあります。

基本問題 - 問題1

問題1 (基本的な統計量の計算)

データベクトル $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$ について、以下を計算せよ。

a) 平均 $\bar{x}$

b) 偏差ベクトル $\mathbf{d}$

c) 分散 σ^2

d) 標準偏差 σ

基本問題 - 問題2

問題2 (ベクトル表記を用いた計算)

データベクトル $\mathbf{y} = \begin{pmatrix} 10 \\ 20 \\ 30 \\ 40 \end{pmatrix}$ と、全ての要素が1である4次元ベクトル $\mathbf{1}_4$ を用いて、以下を計算せよ。

a) 平均 $\bar{y} = \frac{1}{4}\mathbf{1}_4^T\mathbf{y}$

b) 偏差ベクトル $\mathbf{d_y} = \mathbf{y} - \bar{y}\mathbf{1}_4$

c) 分散 $\sigma_y^2 = \frac{1}{4}\mathbf{d_y}^T\mathbf{d_y}$

基本問題 - 問題3

問題3 (分散の別公式の利用)

データセット $\{2, 4, 9\}$ について、分散の別公式

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

を用いて分散を計算せよ。

基本問題 - 問題4 (1/2)

問題4 (標準化を含む総合問題)

ある少人数のグループの小テストの得点（単位：点）が以下のように与えられている。

$$\mathbf{s} = \begin{pmatrix} 6 \\ 8 \\ 10 \end{pmatrix}$$

a) 平均得点 $\bar{s}$ を計算せよ。

b) 偏差ベクトル $\mathbf{d_s}$ を求めよ。

c) 分散 σ_s^2 と標準偏差 σ_s を計算せよ。

d) 各学生の得点を標準化（zスコア化）せよ。標準化された値 z_i は $z_i = \frac{s_i - \bar{s}}{\sigma_s}$ で定義される。

e) 標準化されたデータの平均と分散を計算し、それぞれ0と1になることを確認せよ（計算誤差も考慮すること）

基本問題 - 問題5 (1/2)

問題5 (データの変換と統計量の変化)

データベクトル $\mathbf{u} = \begin{pmatrix} 0 \\ 0 \\ 10 \\ 10 \end{pmatrix}$ について、以下を考察し、計算せよ。

a) 平均 $\bar{u}$ と 分散 σ_u^2 を計算せよ。

b) 各データ点に定数 $c = 5$ を加えた新しいデータベクトル $\mathbf{v} = \mathbf{u} + c\mathbf{1}_4$ を考える。$\mathbf{v}$ の平均 $\bar{v}$ と分散 σ_v^2 を計算し、$\bar{u}, \sigma_u^2$ と比較せよ。

c) 各データ点を定数 $k = 2$ 倍した新しいデータベクトル $\mathbf{w} = k\mathbf{u}$ を考える。$\mathbf{w}$ の平均 $\bar{w}$ と分散 σ_w^2 を計算し、$\bar{u}, \sigma_u^2$ と比較せよ。

(ヒント：一般に、元のデータの平均を μ、分散を σ^2 とすると、データ全体に定数 c を加えたデータの平均は $\mu + c$、分散は σ^2。データ全体を定数 k 倍したデータの平均は $k\mu$、分散は $k^2\sigma^2$ となる。)

4. まとめと次回予告 (1/2)

本講義の重要なポイント：

1. **データベクトルの導入**: 記述の簡潔化、計算効率化、多次元への拡張。
2. **1ベクトルの活用**: 総和や平均計算に役立つ。
3. **平均値**: $\bar{x} = \frac{1}{n}(\mathbf{1}_n^T \mathbf{x})$
4. **偏差ベクトル**: $\mathbf{d} = \mathbf{x} - \bar{x}\mathbf{1}_n$ (総和 $\mathbf{1}_n^T \mathbf{d} = 0$)
5. **分散**: $\sigma^2 = \frac{1}{n}(\mathbf{d}^T \mathbf{d}) = \left(\frac{1}{n}\mathbf{x}^T \mathbf{x}\right) - \bar{x}^2$
6. **標準偏差**: $\sigma = \sqrt{\sigma^2}$ (元のデータと同じ単位)
7. **偏差値**

$$T_i = 50 + 10 \times \frac{x_i - \bar{x}}{\sigma}$$

5. まとめと次回予告 (2/2)

これらのベクトルを用いたデータの扱いや統計量の概念は、今後の機械学習やより複雑なデータ分析手法（多変量データ分析、線形回帰モデル、主成分分析など）を学ぶ上での非常に重要な基礎となります。

次回の講義予告

今回の内容をさらに発展させ、1次元データだけでなく、複数の特徴量を持つ「2次元データ」（あるいはより一般の多次元データ）を扱います。
具体的には、そのようなデータを「行列」を用いて表現する方法や、複数の変数間の関係性を示す「共分散」などについて学んでいく予定です。