Data Representation using Vectors and Matrices

Dr. Shailesh Sivan

June 10, 2025

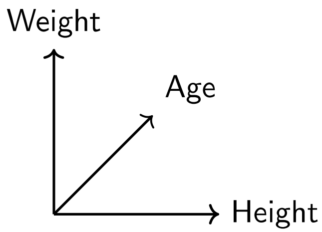
Introduction

- Data in computer systems is structured numerically.
- Linear algebra offers a compact and efficient framework to represent and analyze data.
- Key representations:
  - **Vectors:** One-dimensional arrays representing data points or features.
  - **Matrices:** Two-dimensional arrays representing datasets.

Vectors: Concept and Notation

- A vector is a list of values arranged in a single column or row.
- Example (feature vector of a student):

$$\mathbf{x} = \begin{bmatrix} \text{Height} \\ \text{Weight} \\ \text{Age} \end{bmatrix} = \begin{bmatrix} 170 \\ 65 \\ 21 \end{bmatrix} \in \mathbb{R}^3$$



Data in $\mathbb{R}^n$: From 1D to nD

Understanding data as points in n-dimensional space:

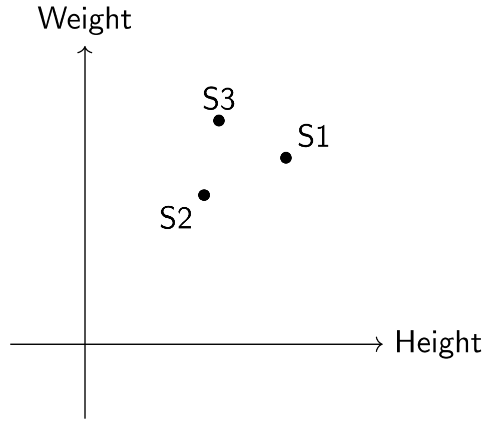
- **1D Example ($\mathbb{R}$):**
  - Single feature, e.g., Temperature.
  - Data: $x = [30], [32], [28] \in \mathbb{R}^1$
- **2D Example ($\mathbb{R}^2$):**
  - Two features, e.g., Height and Weight.
  - Data: $x = \begin{bmatrix} 170 \\ 65 \end{bmatrix}, \begin{bmatrix} 160 \\ 60 \end{bmatrix} \in \mathbb{R}^2$

Data in $\mathbb{R}^n$: From 1D to nD

Understanding data as points in n-dimensional space:

- **3D Example ($\mathbb{R}^3$):**
  - Three features, e.g., Height, Weight, Age.
  - Data: $x = \begin{bmatrix} 170 \\ 65 \\ 21 \end{bmatrix} \in \mathbb{R}^3$
- **nD Example ($\mathbb{R}^n$):**
  - For n features: $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$
  - Used in ML, e.g., each image = 784D vector, BERT embedding = 768D.

2D Feature Space ($\mathbb{R}^2$)



Each point: a 2D data vector $\in \mathbb{R}^2$

Vector Spaces in Data

- Each data point with n features can be viewed as a vector in $\mathbb{R}^n$.
- A dataset with m samples becomes a collection of such vectors:

$$D = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\} \subset \mathbb{R}^n$$

- Here, $\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$ represents the i^{th} sample.
- This view allows us to apply geometric and algebraic operations in n-dimensional space.

Dataset as a Vector Space in $\mathbb{R}^n$

- A collection of vectors in $\mathbb{R}^n$ can span a subspace—a vector space that satisfies:
  - Closure under vector addition.
  - Closure under scalar multiplication.
- This space allows us to analyze patterns and relationships using:
  - **Linear combinations:** New vectors formed using existing ones.
  - **Subspaces:** Lower-dimensional representations of the original data.
- Example of linear combination:

$$\mathbf{y} = \alpha \mathbf{x}_1 + \beta \mathbf{x}_2, \quad \alpha, \beta \in \mathbb{R}$$

  where $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ are feature vectors.
- Such combinations span a **data subspace**, e.g., the plane defined by $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^3$.

Matrices: Concept and Notation

- A matrix organizes data as rows and columns.
- Each row: a data sample. Each column: a feature.
- Matrix $A \in \mathbb{R}^{m \times n}$:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

From Raw Data to Matrix Representation

Raw Tabular Data:

Student	Height	Weight	Age
S1	170	65	21
S2	160	60	22
S3	180	70	20

Matrix Form:

$$A = \begin{bmatrix} 170 & 65 & 21 \\ 160 & 60 & 22 \\ 180 & 70 & 20 \end{bmatrix}$$ (Rows: Students, Columns: Features)

Matrix Visualization

$$\begin{array}{c} \\ \textbf{S1} \\ \textbf{S2} \\ \textbf{S3} \end{array} \begin{array}{c} \begin{array}{ccc} \textbf{Height} & \textbf{Weight} & \textbf{Age} \end{array} \\ \begin{bmatrix} 170 & 65 & 21 \\ 160 & 60 & 22 \\ 180 & 70 & 20 \end{bmatrix} \end{array}$$

Dot Product: Measuring Similarity

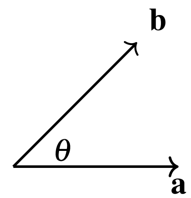
- Dot product measures similarity between vectors:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

- Also expressed as:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos(\theta)$$

- If $\theta = 90^\circ$, vectors are orthogonal.



Basic Vector and Matrix Operations

- **Dot Product:** $\mathbf{a} \cdot \mathbf{b} = \sum a_i b_i$
- **Matrix Multiplication:**

$$C = AB \Rightarrow C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

- **Transpose:** A^T interchanges rows and columns

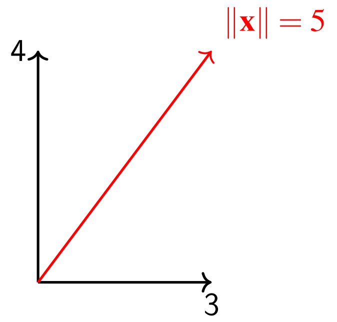
Norms and Magnitude of Vectors

- Norm of a vector (length):

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

- Example:

$$\| \begin{bmatrix} 3 \\ 4 \end{bmatrix} \| = \sqrt{3^2 + 4^2} = 5$$



Applications in Data Science

- **Machine Learning:** Features and weights as vectors.
- **Image Processing:** Pixels arranged in matrices.
- **Recommendation Systems:** Sparse user-item matrices.
- **NLP:** Word embeddings are high-dimensional vectors.

Summary

- Vectors represent individual features or data samples.
- Matrices structure complete datasets.
- Linear algebra operations are central to data modeling.
- Foundation for advanced techniques like PCA, SVD, and ML.