Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Thang Vu

Math for Machine Learning

Introduction

Contents

1 Introduction

2 Vectors

3 Matrices

4 Linear Independence and Basis

5 Linear Mappings

Linear Algebra for CL

Words, sentences, and documents can be represented as vectors or matrices

Basis for measuring similarity, relationships, and patterns in language

Enables quantitative analysis of linguistic entities

Scalars, Vectors, and Matrices

In the world of numbers, think of:

- Scalars as individual numbers
- Vectors as lists of numbers
- Matrices as tables of numbers

Examples:

- Scalar as a sentiment score
- Vector can represent a word
- Matrices can represent a sentence

Vectors

2

Vectors

- You can not compare apples and oranges!
- Idea: But we could keep track of both types of fruits with a single object, in the form of a tuple which we call a vector:

$$\begin{bmatrix} \#\text{apples} \\ \#\text{oranges} \end{bmatrix}$$

- Consider having a fruit bowl with 1 apple and 2 oranges
- Let's call the vector that holds both these numbers at the same time $\mathbf{v}$:

$$\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Vector Operations: Definition

Addition

For two vectors $v, w \in \mathbb{R}^n$, addition is performed by adding together elements of the same index:

$$\mathbf{v} + \mathbf{w} = \begin{bmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_n \end{bmatrix} = \begin{bmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \dots v_n + w_n \end{bmatrix} \in \mathbb{R}^n$$

Multiplication with a scalar

For a vector $v \in \mathbb{R}^n$ and a scalar $\lambda \in \mathbb{R}$, scalar multiplication is performed by element-wise multiplication with that scalar:

$$\lambda \mathbf{v} = \lambda \cdot \begin{bmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{bmatrix} = \begin{bmatrix} \lambda \cdot v_1 \\ \lambda \cdot v_2 \\ \dots \\ \lambda \cdot v_n \end{bmatrix} \in \mathbb{R}^n$$

Vector Operations

Addition

A friend now brings us another bowl of fruit $\mathbf{w}$ with another 2 apples and 4 oranges: $\mathbf{v} + \mathbf{w} = \mathbf{u}$

Example: $\mathbf{u} = \mathbf{v} + \mathbf{w} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$

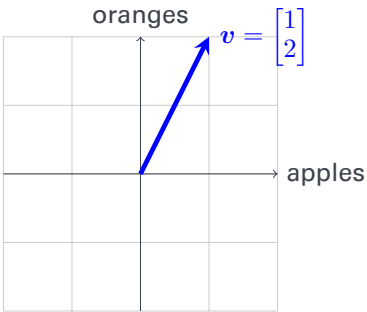
Multiplication with a scalar

We can also multiply vectors by a scalar $\lambda \in \mathbb{R}$, e.g. express our updated fruit bowl contents as $\lambda \cdot \mathbf{v} = \mathbf{u}$

Example: $\mathbf{u} = 3\mathbf{v} = 3 * \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$

Vectors: Geometric Interpretation

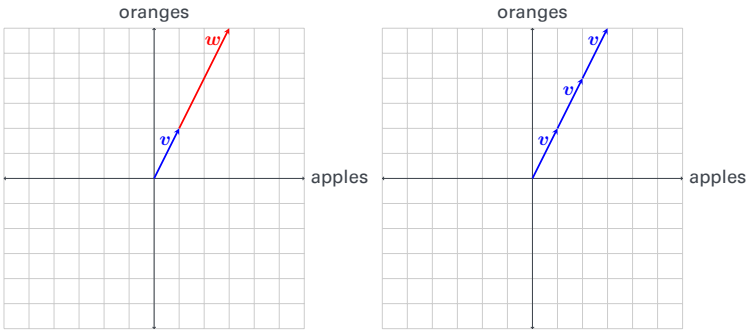
- We can also interpret vectors gemoetrically, with a direction and magnitude:



Vectors: Geometric Interpretation

- ... and also visualize the addition / scaling operations:

$$\mathbf{v} + \mathbf{w} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix} = 3\mathbf{v}:$$



Vectors and Vector Spaces

- Vectors are elements of a vector space
- Our example vectors so far had two components (two real numbers)
  - We can refer to individual components using indices (e.g., apples: v_1, oranges: v_2)
  - Both v_1, v_2 are real numbers: $v_1, v_2 \in \mathbb{R}$
  - $\rightarrow$ $\mathbf{v}$ is an element of the 2-dimensional vector space $\mathbb{R} \times \mathbb{R}$, or simply $\mathbb{R}^2$: $\mathbf{v} \in \mathbb{R}^2$

Definition (Vector Space)

A real-valued vector space $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with two operations, addition and multiplication:

- $+ : \mathcal{V} \times \mathcal{V} \rightarrow \mathcal{V}$ (inner operation)
- $\cdot \ : \mathbb{R} \times \mathcal{V} \rightarrow \mathcal{V}$ (outer operation)

Vectors and Vector Spaces: Examples

- In NLP, we can use vectors to represent words in a numerical form (so that we can perform computations with them)

  Example: One-hot encoding

  - A vector $\mathbf{v} \in \{0,1\}^n$, where n is the size of the vocabulary
  - All entries of a vector are 0, except one that is 1
  - In this way, we can map each possible word uniquely to a vector (each word corresponds to a different component being 1)
  - Example:
    - Vocabulary: V = {NLP, is, fun}

$$v_{\text{NLP}} = \begin{bmatrix}1\\0\\0\end{bmatrix}, v_{\text{is}} = \begin{bmatrix}0\\1\\0\end{bmatrix}, v_{\text{fun}} = \begin{bmatrix}0\\0\\1\end{bmatrix}$$

Vectors and Vector Spaces: Examples

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/OD3Y`

Vector Spaces in NLP

Word Embeddings
High-dimensional vectors that represent words or phrases in a continuous vector space, e.g. Word2Vec or GloVe. They capture semantic relationships between words.

Semantic Similarity
Cosine similarity is often used to measure the similarity between two vectors in the space, indicating how similar two pieces of text are in terms of their meaning.

Dimensionality Reduction
Techniques like PCA or t-SNE can reduce the dimensionality of text data for visualization or further processing.

Vector Spaces in NLP

Challenges

- High dimensionality: The ”curse of dimensionality” can make some algorithms run slowly.
- Sparsity: Many text representations (like one-hot encoding) can be very sparse, meaning they have a lot of zeros.
- One word can have multiple meanings (e.g., ”bank” as a financial institution and ”bank” of a river).

Subword Embeddings

Representing words based on smaller chunks or characters can help in understanding morphologically rich languages or out-of-vocabulary words.

Vector Space Example: Word Embeddings

Figure: A two-dimensional (t-SNE) projection of embeddings for some words and phrases. Trained for sentiment analysis.

Vector Spaces: Properties of Operations

- Distributive
  - $\forall \lambda \in \mathbb{R}, \mathbf{v}, \mathbf{w} \in \mathcal{V} : \lambda \cdot (\mathbf{v} + \mathbf{w}) = \lambda \mathbf{v} + \lambda \mathbf{w}$
  - $\forall \lambda, \psi \in \mathbb{R}, \mathbf{v} \in \mathcal{V} : (\lambda + \psi) \cdot \mathbf{v} = \lambda \cdot \mathbf{v} + \psi \cdot \mathbf{v}$
- Associative
  - $\forall \lambda, \psi \in \mathbb{R}, \mathbf{v} \in \mathcal{V} : \lambda \cdot (\psi \cdot \mathbf{v}) = (\lambda \cdot \psi) \cdot \mathbf{x}$
- Neutral element w.r.t. outer operation
  - $\forall \mathbf{v} \in \mathcal{V} : 1 \cdot \mathbf{v} = \mathbf{v}$

Vector spaces we have seen so far: $\mathcal{V} = \mathbb{R}^n$, $n \in \mathbb{N}$.

Element-wise Product for Vectors

Definition (Hadamard Product)

Given two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, the Hadamard product, denoted by $\odot$, is defined as the element-wise product of the vectors.

Example:

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\mathbf{a} \odot \mathbf{b} = \begin{bmatrix} a_1 \cdot b_1 \\ a_2 \cdot b_2 \\ \vdots \\ a_n \cdot b_n \end{bmatrix}$$

Element-wise Product for Vectors

Example:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 4 \end{bmatrix}$$

$$\mathbf{x} \odot \mathbf{y} = \begin{bmatrix} 1 \cdot 1 \\ 0 \cdot 1 \\ 0.5 \cdot 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$$

Element-wise Product for Vectors

Example usages of the Hadamard product in NLP:

- Gating mechanisms of LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) cells
- Fusion techniques, e.g. in multi-modal NLP tasks such as VQA to combine the representations of the different modalities

Inner Products

Definition (Scalar-Product)

An inner product of $\mathcal{V}$ is defined as:

$$\Omega : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$$

Significance in ML & CL:
The inner product indicates similarity between vectors.

Inner Products

Definition (Inner Product)

The properties of an inner product are defined as:

Linearity:

$$\Omega(\alpha x + \beta y, z) = \alpha\Omega(x, z) + \beta\Omega(y, z)$$
$$\Omega(x, \alpha y + \beta z) = \alpha\Omega(x, y) + \beta\Omega(x, z)$$

Symmetry:

$$\forall x, y \in \mathcal{V} : \Omega(x, y) = \Omega(y, x)$$

Positive:

$$\Omega(x, x) = 0 \iff x = 0$$
$$\forall x \in \mathcal{V} \backslash \{0\} : \Omega(x, x) > 0$$

Scalar/Dot-Product

Definition (Scalar-Product)

A scalar product or dot product of $\mathcal{V}$ is a particular type of an inner product and is defined as:

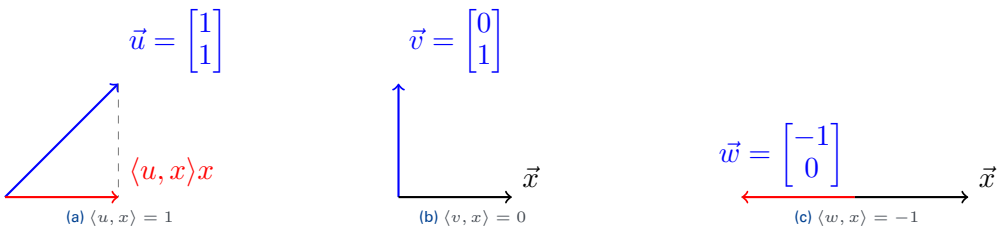
$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$$

Significance in ML & CL

Dot product indicates similarity: if it's positive, vectors point in a similar direction; if zero, vectors are orthogonal; if negative, vectors point in opposite directions.

Scalar/Dot-Product: Geometric Interpretation

- Scalar products compute the magnitude of a vector (here: $\mathbf{v}$) in the direction of another vector (here: a vector $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$)



Scalar/Dot-Product

Example

$$x = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \; y = \begin{bmatrix} 1 \\ 1 \\ 4 \end{bmatrix}$$

$$\langle x, y \rangle = ?$$

Scalar/Dot-Product

Example

$$x = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \; y = \begin{bmatrix} 1 \\ 1 \\ 4 \end{bmatrix}$$

$$\langle x, y \rangle = 9$$

Vector Norms

The norm (or magnitude) of a vector indicates its length.

General L_p norm:

$$||x||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

Norms can help normalize vector representations.

L_1 and L_2 Norms

L_1 Norm (Manhattan or Taxicab norm)

$$||x||_1 = \sum_{i=1}^{n} |x_i|$$

Intuition: Distance in a grid-based path. Used for inducing sparsity in regularization.

L_2 Norm (Euclidean norm)

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Intuition: Straight-line distance in Euclidean space. Used in clustering and distance-based algorithms.

L_2: Example

For L_2 norm (Euclidean norm):

$$||a||_2 = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$$

Example using L_2 norm:

$$|| \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} ||_2 = \sqrt{1 + 4 + 4} = 3$$

Norms in Machine Learning

Regularization

- L_1 Norm (Lasso): Encourages model sparsity.
- L_2 Norm (Ridge): Penalizes large coefficients without enforcing sparsity.

Neural Networks

- Weight Decay: L_2 norm prevents overfitting by penalizing large weights.

Difference Between Norm and Distance

Norm

Measures the magnitude of a single vector.

For vector v: $||v||$

Example: $||\begin{bmatrix}1\\2\\2\end{bmatrix}||$ measures the length of this vector.

Distance

Measures the separation between two vectors or points.

For vectors u and v: $d(u, v)$

Example: $d(\begin{bmatrix}1\\0\\2\end{bmatrix}, \begin{bmatrix}1\\1\\4\end{bmatrix})$ measures the separation between these vectors.

Connection: The distance between vectors u and v can be defined using the norm as $||u - v||$.

Cosine Similarity

Definition (Cosine Similarity)

The cosine similarity measures the cosine of the angle between two non-zero vectors. Given two vectors $\mathbf{x}$ and $\mathbf{y}$, their cosine similarity is defined as:

$$\cos(\theta) = \frac{\langle x, y \rangle}{||\mathbf{x}|| \times ||\mathbf{y}||}$$

where $\langle x, y \rangle$ is the dot product, and $||\mathbf{x}||$ and $||\mathbf{y}||$ are the magnitudes (L2 norms) of the vectors.

Properties

- The value ranges from -1 (completely dissimilar) to 1 (identical).
- A value of 0 indicates orthogonality or decorrelation.
- Often used in NLP to measure semantic similarity between texts.

Cosine Similarity

Example

If $\mathbf{x} = [1, 2]$ and $\mathbf{y} = [2, 3]$, then:

$$\cos(\theta) = \frac{1 \times 2 + 2 \times 3}{\sqrt{1^2 + 2^2} \times \sqrt{2^2 + 3^2}}$$

Cosine similarity: Examples

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/OD3Y`

Matrices

3

Definition of Matrices

Can be used to compactly represent:

- a 2D input
- linear functions (linear mappings)

Definition (Matrix)

With $m, n \in \mathbb{N}$ a real-valued (m, n) matrix A is an $m \times n$-tuple of elements $a_{ij}, i = 1, \ldots, m, j = 1, \ldots, n$, which is ordered:

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}, a_{i,j} \in \mathbb{R}$$

Matrix Addition (1)

Definition (Matrix Addition)

The sum of two matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{m \times n}$ is defined as the element- wise sum:

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Matrix Addition (2)

Example

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \\ -1 & 5 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 3}, \quad \mathbf{B} = \begin{bmatrix} 1 & 3 & 2 \\ -2 & 5 & 2 \\ 0 & 3 & 7 \end{bmatrix} \in \mathbb{R}^{3\times 3}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \\ -1 & 5 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 3 & 2 \\ -2 & 5 & 2 \\ 0 & 3 & 7 \end{bmatrix} = \mathbf{?}$$

Matrix Addition (3)

Example

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \\ -1 & 5 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 3}, \quad \mathbf{B} = \begin{bmatrix} 1 & 3 & 2 \\ -2 & 5 & 2 \\ 0 & 3 & 7 \end{bmatrix} \in \mathbb{R}^{3\times 3}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \\ -1 & 5 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 3 & 2 \\ -2 & 5 & 2 \\ 0 & 3 & 7 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 4 & 6 \\ -1 & 8 & 10 \end{bmatrix}$$

Matrix Multiplication (1)

Matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times k}$, the elements c_{ij} of the product $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times k}$ are computed as:

Definition (Matrix Multiplication)

$$c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj}, \quad i = 1, \ldots, m \quad j = 1, \ldots, k$$

To compute the element c_{ij} we multiply the ith row of matrix $\mathbf{A}$ with the jth column of matrix $\mathbf{B}$, and sum them up.

This is called dot-product, scalar-product, or inner-product of the corresponding row and column.

Matrix Multiplication (2)

Matrices can only be multiplied if their "neighboring" dimensions match

A $n \times k$ matrix can only be multiplied with a $k \times m$ matrix, and only from the left side.

Matrix multiplication is not defined as element-wise multiplication.

- this is called Hadamard-Product
- matrices need to be of the same shape as for the matrix addition

Matrix Multiplication (3)

Example

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \end{bmatrix} \in \mathbb{R}^{2\times 3}, \quad \mathbf{B} = \begin{bmatrix} 1 & 3 \\ -2 & 5 \\ 0 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

$$\mathbf{A} * \mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ -2 & 5 \\ 0 & 3 \end{bmatrix} = \textbf{?}$$

$$\mathbf{B} * \mathbf{A} = \begin{bmatrix} 1 & 3 \\ -2 & 5 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \end{bmatrix} = \textbf{?}$$

Matrix Multiplication (4)

Example

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \end{bmatrix} \in \mathbb{R}^{2\times 3}, \quad \mathbf{B} = \begin{bmatrix} 1 & 3 \\ -2 & 5 \\ 0 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

$$\mathbf{A} * \mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ -2 & 5 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 11 \\ 5 & 16 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

$$\mathbf{B} * \mathbf{A} = \begin{bmatrix} 1 & 3 \\ -2 & 5 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \end{bmatrix} = \begin{bmatrix} 11 & -2 & 12 \\ 11 & -7 & 20 \\ 9 & -3 & 12 \end{bmatrix} \in \mathbb{R}^{3\times 3}$$

$\Rightarrow$ matrix multiplication is not commutative: $\mathbf{BA} \neq \mathbf{AB}$

Matrix Scalar Multiplication (1)

A scalar multiplication of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, a scalar$\lambda \in \mathbb{R}$ is defined as a element-wise multiplication:

Definition (Matrix Scalar Multiplication)

$$A * \lambda = \begin{bmatrix} a_{11} * \lambda & a_{12} * \lambda & \dots & a_{1n} * \lambda \\ a_{21} * \lambda & a_{22} * \lambda & \dots & a_{2n} * \lambda \\ \vdots & \vdots & & \vdots \\ a_{m1} * \lambda & a_{m2} * \lambda & \dots & a_{mn} * \lambda \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Matrix Scalar Multiplication (2)

Example

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \\ -1 & 5 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 3}, \quad \lambda = 2$$

$$\mathbf{A} * \lambda = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \\ -1 & 5 & 3 \end{bmatrix} * 2 = \mathbf{?}$$

Matrix Scalar Multiplication (3)

Example

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \\ -1 & 5 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 3}, \quad \lambda = 2$$

$$\mathbf{A} * \lambda = \begin{bmatrix} 2 & 1 & 0 \\ 3 & -1 & 4 \\ -1 & 5 & 3 \end{bmatrix} * 2 = \begin{bmatrix} 4 & 2 & 0 \\ 6 & -2 & 8 \\ -2 & 10 & 6 \end{bmatrix}$$

Important Matrices

Identity Matrix

Inverse Matrix

Transpose Matrix

Identity Matrix

Definition (Identity Matrix)

In $\mathbb{R}^{n \times n}$ it is defined as:

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{bmatrix}$$

A $n \times n$ matrix that contains ones on the diagonal and zero everywhere else.

Inverse Matrix

Definition (Inverse Matrix)

Matrix $\mathbf{A}$ is called invertible if there exists a matrix $\mathbf{A}^{-1}$, such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n = \mathbf{A}^{-1}\mathbf{A}$

Example

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -7 & -7 & 6 \\ 2 & 1 & -1 \\ 4 & 5 & -4 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix} \begin{bmatrix} -7 & -7 & 6 \\ 2 & 1 & -1 \\ 4 & 5 & -4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -7 & -7 & 6 \\ 2 & 1 & -1 \\ 4 & 5 & -4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix}$$

$$\mathbf{A}\mathbf{B} = \mathbf{I} = \mathbf{B}\mathbf{A}$$

Transpose

Definition (Transpose)

For $\mathbf{A} \in \mathbb{R}^{m \times n}$ the matrix $\mathbf{B} \in \mathbb{R}^{n \times m}$ with $b_{ij} = a_{ji}$ is called the transpose of $\mathbf{A}$. We write $\mathbf{B} = \mathbf{A}^T$.

$\mathbf{A}^T$ can be obtained by switching rows with columns of $\mathbf{A}$.

Example

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix}$, then $\mathbf{A}^T = \begin{bmatrix} 1 & 4 & 6 \\ 2 & 4 & 7 \\ 1 & 5 & 7 \end{bmatrix}$

Matrices in NLP

Examples

- Trainable parameters of neural networks are matrices
- Word embeddings, where each word is represented as a vector, the whole sentence as a matrix in sentence classification tasks
- Adjacency matrix, which contains the edge information of a graph, e.g. dependency graphs in NLP

Live Voting

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/OD3Y`

Linear Independence and Basis

4

Linear Combination

Definition (Linear Combination)

Consider a vector space $\mathcal{V}$ and a finite number of vectors $x_1, x_2, \ldots, x_k \in \mathcal{V}$. Then every $v \in \mathcal{V}$ of the form

$$v = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_k x_k = \sum_{i=1}^{k} \lambda_i x_i \in \mathcal{V}$$

with $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$is a linear combination of the vectors $x_1, \ldots, x_k$.

Linear Independence

The zero vector can also be written as a linear combination, because $0 = \sum_{i=1}^{k} 0x_i$ is always true.

Definition (Linear Dependence)

If there exists a non-trivial solution to generate $0 = \sum_{i=1}^{k} \lambda_i x_i$ with at least one $\lambda_i \neq 0$ then the vectors $x_1, \ldots, x_k$ are linearly dependent.

Definition (Linear Independence)

If only the trivial solution exists, i.e. $\lambda_1 = \ldots = \lambda_i = 0$, then the vectors $x_1, \ldots, x_k$ are linearly independent.

Intuition of Linear Independence

Vectors being linearly independent means that no vector in the set can be written as a linear combination of the others.

Intuitively, a set of linearly independent vectors consists of vectors that have no redundancy, i.e. if we remove any of those vectors from the set, we will loose something.

In $\mathbb{R}^2$:

- Two vectors are linearly dependent if they lie on the same line or are parallel to each other.
- Two vectors are linearly independent if they are not parallel and do not lie on the same line.

Basis

Definition (Basis)

The basis vectors of a vector space $\mathcal{V}$ is the minimal set of of vectors $\mathcal{A}$ that can generate any vector in $\mathcal{V}$ as a linear combination.

- Canonical/standard base in $\mathbb{R}^3$

$$\mathcal{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Different bases in $\mathbb{R}^3$

$$\mathcal{B}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \mathcal{B}_2 = \begin{bmatrix} 0.5 & 1.8 & -2.2 \\ 0.8 & 0.3 & -1.3 \\ 0.4 & 0.3 & 3.5 \end{bmatrix}$$

Live Voting

Figure: Link to Ilias live voting: `https://ilias3.uni-stuttgart.de/vote/OD3Y`

Linear Mappings

5

Linear Mappings (Functions)

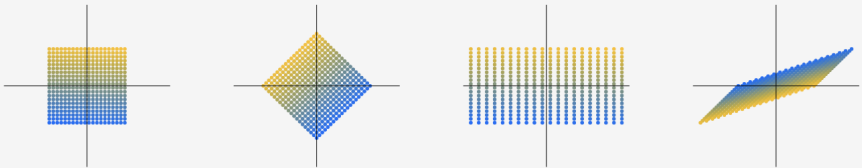
Definition (Linear Mappings)

Given two vector spaces $\mathbf{V}, \mathbf{W}$, a linear mapping $\Phi : \mathbf{V} \to \mathbf{W}$
$\forall x, y \in \mathbf{V}, \forall \lambda, \psi \in \mathbb{R} : \Phi(\lambda x + \psi y) = \lambda \Phi(x) + \psi \Phi(y)$

Examples

- $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a linear regression model that maps an input signal with n dimensions to an output scalar
- A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ applied to $\mathbf{x} \in \mathbb{R}^n$: $\mathbf{Ax} = \mathbf{y} \in \mathbb{R}^m$ is a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$

Examples



(a) Original data. (b) Rotation by 45°. (c) Stretch along the horizontal axis. (d) General linear mapping.

We consider three linear transformations of a set of vectors in $\mathbb{R}^2$ with the transformation matrices

$$\boldsymbol{A}_1 = \begin{bmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{A}_3 = \frac{1}{2}\begin{bmatrix} 3 & -1 \\ 1 & -1 \end{bmatrix}. \quad (2.97)$$

Figure: Examples of linear transformations as matrices. Figure taken from [1].

Properties of Linear Mappings

A function $\Phi : \mathbf{V} \to \mathbf{W}$ is a linear mapping if and only if:

- $\Phi(\mathbf{u} + \mathbf{v}) = \Phi(\mathbf{u}) + \Phi(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in \mathbf{V}$.
- $\Phi(c\mathbf{v}) = c\Phi(\mathbf{v})$ for all scalars c and all vectors $\mathbf{v} \in \mathbf{V}$.

These properties mean linear mappings preserve both vector addition and scalar multiplication.

References

[1] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, Mathematics for Machine Learning. Cambridge University Press, 2020.