MADS-MMS – Mathematics and Multivariate Statistics

Representing Tabular Data

Prof. Dr. Stephan Doerfel

FACHHOCHSCHULE KIEL
University of Applied Sciences

Moodle (WiSe 24/25)

Agenda

Motivation

Data

Vectors

Outline

Motivation

- When modeling data science problems there are two main aspects to consider:

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects
  - operations and comparisons of data objects

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects
  - operations and comparisons of data objects
- Often, data and operations can be modeled as vector space with vector operations

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects
  - operations and comparisons of data objects
- Often, data and operations can be modeled as vector space with vector operations
  - Geometric Vectors (arrows in a coordinate system)

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects
  - operations and comparisons of data objects
- Often, data and operations can be modeled as vector space with vector operations
  - Geometric Vectors (arrows in a coordinate system)
  - Polynomials

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects
  - operations and comparisons of data objects
- Often, data and operations can be modeled as vector space with vector operations
  - Geometric Vectors (arrows in a coordinate system)
  - Polynomials
  - The "Bag of Words" model, models texts as vectors, where each entry represents to frequency of a word, even though, all texts together do not yield a vector space

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects
  - operations and comparisons of data objects
- Often, data and operations can be modeled as vector space with vector operations
  - Geometric Vectors (arrows in a coordinate system)
  - Polynomials
  - The "Bag of Words" model, models texts as vectors, where each entry represents to frequency of a word, even though, all texts together do not yield a vector space
  - Audio Signals of the same length can be modeled as a vector space (entries are sound pressure measurements)

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects
  - operations and comparisons of data objects
- Often, data and operations can be modeled as vector space with vector operations
  - Geometric Vectors (arrows in a coordinate system)
  - Polynomials
  - The "Bag of Words" model, models texts as vectors, where each entry represents to frequency of a word, even though, all texts together do not yield a vector space
  - Audio Signals of the same length can be modeled as a vector space (entries are sound pressure measurements)
  - $\mathbb{R}^n$ (focus of this module)

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects
  - operations and comparisons of data objects
- Often, data and operations can be modeled as vector space with vector operations
  - Geometric Vectors (arrows in a coordinate system)
  - Polynomials
  - The "Bag of Words" model, models texts as vectors, where each entry represents to frequency of a word, even though, all texts together do not yield a vector space
  - Audio Signals of the same length can be modeled as a vector space (entries are sound pressure measurements)
  - $\mathbb{R}^n$ (focus of this module)
- Linear algebra abstracts the similarities of those examples

Motivation

- When modeling data science problems there are two main aspects to consider:
  - data objects
  - operations and comparisons of data objects
- Often, data and operations can be modeled as vector space with vector operations
  - Geometric Vectors (arrows in a coordinate system)
  - Polynomials
  - The "Bag of Words" model, models texts as vectors, where each entry represents to frequency of a word, even though, all texts together do not yield a vector space
  - Audio Signals of the same length can be modeled as a vector space (entries are sound pressure measurements)
  - $\mathbb{R}^n$ (focus of this module)
- Linear algebra abstracts the similarities of those examples
- Analytical geometry helps modeling notions like distances or angles in abstract spaces

Chapter Goals

- understand measurement levels of data

Chapter Goals

- understand measurement levels of data
- mathematical modeling of tabular data

Chapter Goals

- understand measurement levels of data
- mathematical modeling of tabular data
- basics of vectors and matrices

Chapter Goals

- understand measurement levels of data
- mathematical modeling of tabular data
- basics of vectors and matrices
- preparation for clustering, SVMs, neural nets, ...

Outline

- Motivation
- Data
- Vectors

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: "food allergy"; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: "likes death metal".

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: “food allergy”; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: “likes death metal”.

Issues regarding this data:

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: "food allergy"; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: "likes death metal".

Issues regarding this data:

- How is the data structured?

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: “food allergy”; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: “likes death metal”.

Issues regarding this data:

- How is the data structured?
- Do we understand the features and their values?

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: "food allergy"; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: "likes death metal".

Issues regarding this data:

- How is the data structured?
- Do we understand the features and their values?
- Which features are relevant for our purpose?

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: "food allergy"; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: "likes death metal".

Issues regarding this data:

- How is the data structured?
- Do we understand the features and their values?
- Which features are relevant for our purpose?
- Which features should we exclude (e.g., for ethical reasons)?

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: “food allergy”; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: “likes death metal”.

Issues regarding this data:

- How is the data structured?
- Do we understand the features and their values?
- Which features are relevant for our purpose?
- Which features should we exclude (e.g., for ethical reasons)?
- Are their missing features, what do they mean?

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: "food allergy"; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: "likes death metal".

Issues regarding this data:

- How is the data structured?
- Do we understand the features and their values?
- Which features are relevant for our purpose?
- Which features should we exclude (e.g., for ethical reasons)?
- Are their missing features, what do they mean?
- Are the values compatible (use same units, same spelling)?

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: "food allergy"; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: "likes death metal".

Issues regarding this data:

- How is the data structured?
- Do we understand the features and their values?
- Which features are relevant for our purpose?
- Which features should we exclude (e.g., for ethical reasons)?
- Are their missing features, what do they mean?
- Are the values compatible (use same units, same spelling)?
- . . . , compliance, representativity, volume, . . .

Data Representation Example 1/2

Example: Alice has height: 1.85, gender: female, MBTI: INTJ, age: 32y, eye-color: blue, education: M. Sc., children: 2, comment: “food allergy”; Bob has height: 179, gender: m, MBTI: ESFP, age: 56y, eye-color: green, education: High School, comment: “likes death metal”.

Issues regarding this data:

- How is the data structured?
- Do we understand the features and their values?
- Which features are relevant for our purpose?
- Which features should we exclude (e.g., for ethical reasons)?
- Are their missing features, what do they mean?
- Are the values compatible (use same units, same spelling)?
- . . . , compliance, representativity, volume, . . .
- How to represent the data?

Data Representation Example 2/2

To address these issues, among others:

- ask domain experts (! a lot!)
- transform the data
- impute missing data
- drop features
- drop data points
- recollect or supplement data
- find a representation that suits the data AND the methods you are going to apply to it

Look at the data (plots, distributions, . . .)!

Data Representation Example 2/2

To address these issues, among others:

- ask domain experts (! a lot!)
- transform the data
- impute missing data
- drop features
- drop data points
- recollect or supplement data
- find a representation that suits the data AND the methods you are going to apply to it

Look at the data (plots, distributions, ...)!

Question the data and your and others assumptions about it!

Tabular Data – Example

In the example, we are given semi-structured data. For the following **tabular representation**, we

- use the name as an ID (will not be used as a feature)

Tabular Data – Example

In the example, we are given semi-structured data. For the following **tabular representation**, we

- use the name as an ID (will not be used as a feature)
- harmonize data – transform *m* into *cm*

Tabular Data – Example

In the example, we are given semi-structured data. For the following **tabular representation**, we

- use the name as an ID (will not be used as a feature)
- harmonize data – transform *m* into *cm*
- extract data – MBTI score to extroversion/introversion

Tabular Data – Example

In the example, we are given semi-structured data. For the following **tabular representation**, we

- use the name as an ID (will not be used as a feature)
- harmonize data – transform *m* into *cm*
- extract data – MBTI score to extroversion/introversion
- drop data – gender (ethical), children (data is missing), comment (unstructured textual data)

Tabular Data – Example

In the example, we are given semi-structured data. For the following **tabular representation**, we

- use the name as an ID (will not be used as a feature)
- harmonize data – transform *m* into *cm*
- extract data – MBTI score to extroversion/introversion
- drop data – gender (ethical), children (data is missing), comment (unstructured textual data)

name	height	ext/int	age	eye-color	education
Alice	185	I	32	blue	M. Sc.
Bob	179	E	56	green	High School

Tabular Data – Example

In the example, we are given semi-structured data. For the following **tabular representation**, we

- use the name as an ID (will not be used as a feature)
- harmonize data – transform *m* into *cm*
- extract data – MBTI score to extroversion/introversion
- drop data – gender (ethical), children (data is missing), comment (unstructured textual data)

name	height	ext/int	age	eye-color	education
Alice	185	I	32	blue	M. Sc.
Bob	179	E	56	green	High School

All the above steps are OUR choices. Different choices might yield different outcome.

Measurement Levels

Features are of different types, so called **measurement levels**:

Measurement Levels

Features are of different types, so called **measurement levels**:

categorical a variable can take one value out of a fix selection of possible values – example: smoker: yes/no

Measurement Levels

Features are of different types, so called **measurement levels**:

categorical a variable can take one value out of a fix selection of possible values – example: smoker: yes/no

ordinal a categorical variable with a linear order on the set of possible values – example: rank in a race

Measurement Levels

Features are of different types, so called **measurement levels**:

categorical a variable can take one value out of a fix selection of possible values – example: smoker: yes/no

ordinal a categorical variable with a linear order on the set of possible values – example: rank in a race

cardinal an ordinal variable, with equal intervals between points – example: points in an exam

Measurement Levels

Features are of different types, so called **measurement levels**:

categorical a variable can take one value out of a fix selection of possible values – example: smoker: yes/no

ordinal a categorical variable with a linear order on the set of possible values – example: rank in a race

cardinal an ordinal variable, with equal intervals between points – example: points in an exam

➔ A variable's level is usually the most specific (highest) level.

Measurement Levels

Features are of different types, so called **measurement levels**:

categorical a variable can take one value out of a fix selection of possible values – example: smoker: yes/no

ordinal a categorical variable with a linear order on the set of possible values – example: rank in a race

cardinal an ordinal variable, with equal intervals between points – example: points in an exam

→ A variable's level is usually the most specific (highest) level.

height	ext/int	age	eye-color	education

Measurement Levels

Features are of different types, so called **measurement levels**:

categorical a variable can take one value out of a fix selection of possible values – example: smoker: yes/no

ordinal a categorical variable with a linear order on the set of possible values – example: rank in a race

cardinal an ordinal variable, with equal intervals between points – example: points in an exam

→ A variable's level is usually the most specific (highest) level.

height	ext/int	age	eye-color	education

Measurement Levels

Features are of different types, so called **measurement levels**:

categorical a variable can take one value out of a fix selection of possible values – example: smoker: yes/no

ordinal a categorical variable with a linear order on the set of possible values – example: rank in a race

cardinal an ordinal variable, with equal intervals between points – example: points in an exam

→ A variable's level is usually the most specific (highest) level.

height	ext/int	age	eye-color	education
card.	cat.	card.	cat./ord.	cat./ord.

Similar features can have different levels. Different data science methods require different levels.

Representations for Data Operations

Each data point can be represented as a row in a table or as a tuple – a list of fix length with one entry for each feature. Example:

Alice (185, I, 32, blue, M. Sc.)

Bob (179, E, 56, green, High School)

Representations for Data Operations

Each data point can be represented as a row in a table or as a tuple – a list of fix length with one entry for each feature. Example:

Alice (185, I, 32, blue, M. Sc.)

Bob (179, E, 56, green, High School)

A key means for data science is to determine distances between instances.

Representations for Data Operations

Each data point can be represented as a row in a table or as a tuple – a list of fix length with one entry for each feature. Example:

Alice (185, I, 32, blue, M. Sc.)

Bob (179, E, 56, green, High School)

A key means for data science is to determine distances between instances.

- E.g. classify similar instances similarly; group similar instances into the same cluster, ...

Representations for Data Operations

Each data point can be represented as a row in a table or as a tuple – a list of fix length with one entry for each feature. Example:

Alice (185, I, 32, blue, M. Sc.)

Bob (179, E, 56, green, High School)

A key means for data science is to determine distances between instances.

- E.g. classify similar instances similarly; group similar instances into the same cluster, ...
- Distance functions exist for data on different measurement levels.

Representations for Data Operations

Each data point can be represented as a row in a table or as a tuple – a list of fix length with one entry for each feature. Example:

Alice (185, I, 32, blue, M. Sc.)

Bob (179, E, 56, green, High School)

A key means for data science is to determine distances between instances.

- E.g. classify similar instances similarly; group similar instances into the same cluster, ...
- Distance functions exist for data on different measurement levels.
- Many distances work in vector spaces.

Outline

Motivation

Data

Vectors

Vector Space

Definition 1 (Real-valued Vector Space)

A **real-valued vector space** $\mathcal{V}$ consists of a set V and operations
$+ : V \times V \to V : (\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$ and
$\cdot : \mathbb{R} \times V \to V : (r, \boldsymbol{v}) \mapsto r \cdot \boldsymbol{v}$
such that the following conditions hold for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $s, t \in \mathbb{R}$:

Vector Space

Definition 1 (Real-valued Vector Space)

A **real-valued vector space** $\mathcal{V}$ consists of a set V and operations
$+ : V \times V \to V : (\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$ and
$\cdot : \mathbb{R} \times V \to V : (r, \boldsymbol{v}) \mapsto r \cdot \boldsymbol{v}$
such that the following conditions hold for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $s, t \in \mathbb{R}$:

1. $\exists 0 \in V : \boldsymbol{v} + 0 = \boldsymbol{v}$ (neutral element of $+$)

Vector Space

Definition 1 (Real-valued Vector Space)

A **real-valued vector space** $\mathcal{V}$ consists of a set V and operations
$+ : V \times V \to V : (\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$ and
$\cdot : \mathbb{R} \times V \to V : (r, \boldsymbol{v}) \mapsto r \cdot \boldsymbol{v}$
such that the following conditions hold for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $s, t \in \mathbb{R}$:

1. $\exists 0 \in V : \boldsymbol{v} + 0 = \boldsymbol{v}$ (neutral element of $+$)
2. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$ (associativity)

Vector Space

Definition 1 (Real-valued Vector Space)

A real-valued vector space $\mathcal{V}$ consists of a set V and operations
$+ : V \times V \to V : (\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$ and
$\cdot : \mathbb{R} \times V \to V : (r, \boldsymbol{v}) \mapsto r \cdot \boldsymbol{v}$
such that the following conditions hold for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $s, t \in \mathbb{R}$:

1. $\exists 0 \in V : \boldsymbol{v} + 0 = \boldsymbol{v}$ (neutral element of $+$)
2. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$ (associativity)
3. $\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{w} + \boldsymbol{v}$ (commutativity)

Vector Space

Definition 1 (Real-valued Vector Space)

A **real-valued vector space** $\mathcal{V}$ consists of a set V and operations
$+ : V \times V \to V : (\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$ and
$\cdot : \mathbb{R} \times V \to V : (r, \boldsymbol{v}) \mapsto r \cdot \boldsymbol{v}$
such that the following conditions hold for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $s, t \in \mathbb{R}$:

1. $\exists 0 \in V : \boldsymbol{v} + 0 = \boldsymbol{v}$ (neutral element of $+$)
2. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$ (associativity)
3. $\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{w} + \boldsymbol{v}$ (commutativity)
4. $\boldsymbol{v} + (-1) \cdot \boldsymbol{v} = 0$ (inverse elements)

Vector Space

Definition 1 (Real-valued Vector Space)

A **real-valued vector space** $\mathcal{V}$ consists of a set V and operations
$+ : V \times V \to V : (\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$ and
$\cdot : \mathbb{R} \times V \to V : (r, \boldsymbol{v}) \mapsto r \cdot \boldsymbol{v}$
such that the following conditions hold for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $s, t \in \mathbb{R}$:

1. $\exists 0 \in V : \boldsymbol{v} + 0 = \boldsymbol{v}$ (neutral element of $+$)
2. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$ (associativity)
3. $\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{w} + \boldsymbol{v}$ (commutativity)
4. $\boldsymbol{v} + (-1) \cdot \boldsymbol{v} = 0$ (inverse elements)
5. $1 \cdot \boldsymbol{v} = \boldsymbol{v}$ (neutral element of $\cdot$)

Vector Space

Definition 1 (Real-valued Vector Space)

A **real-valued vector space** $\mathcal{V}$ consists of a set V and operations
$+ : V \times V \to V : (\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$ and
$\cdot : \mathbb{R} \times V \to V : (r, \boldsymbol{v}) \mapsto r \cdot \boldsymbol{v}$
such that the following conditions hold for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $s, t \in \mathbb{R}$:

1. $\exists 0 \in V : \boldsymbol{v} + 0 = \boldsymbol{v}$ (neutral element of $+$)
2. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$ (associativity)
3. $\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{w} + \boldsymbol{v}$ (commutativity)
4. $\boldsymbol{v} + (-1) \cdot \boldsymbol{v} = 0$ (inverse elements)
5. $1 \cdot \boldsymbol{v} = \boldsymbol{v}$ (neutral element of $\cdot$)
6. $s \cdot (t \cdot \boldsymbol{v}) = (s \cdot t) \cdot \boldsymbol{v}$ (compatibility)

Vector Space

Definition 1 (Real-valued Vector Space)

A real-valued vector space $\mathcal{V}$ consists of a set V and operations
$+ : V \times V \to V : (\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$ and
$\cdot : \mathbb{R} \times V \to V : (r, \boldsymbol{v}) \mapsto r \cdot \boldsymbol{v}$
such that the following conditions hold for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $s, t \in \mathbb{R}$:

1. $\exists 0 \in V : \boldsymbol{v} + 0 = \boldsymbol{v}$ (neutral element of $+$)
2. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$ (associativity)
3. $\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{w} + \boldsymbol{v}$ (commutativity)
4. $\boldsymbol{v} + (-1) \cdot \boldsymbol{v} = 0$ (inverse elements)
5. $1 \cdot \boldsymbol{v} = \boldsymbol{v}$ (neutral element of $\cdot$)
6. $s \cdot (t \cdot \boldsymbol{v}) = (s \cdot t) \cdot \boldsymbol{v}$ (compatibility)
7. $s \cdot (\boldsymbol{v} + \boldsymbol{w}) = s \cdot \boldsymbol{v} + s \cdot \boldsymbol{w}$ (distributivity)

Vector Space

Definition 1 (Real-valued Vector Space)

A real-valued vector space $\mathcal{V}$ consists of a set V and operations
$+ : V \times V \to V : (\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$ and
$\cdot : \mathbb{R} \times V \to V : (r, \boldsymbol{v}) \mapsto r \cdot \boldsymbol{v}$
such that the following conditions hold for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$ and $s, t \in \mathbb{R}$:

1. $\exists 0 \in V : \boldsymbol{v} + 0 = \boldsymbol{v}$ (neutral element of $+$)
2. $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$ (associativity)
3. $\boldsymbol{v} + \boldsymbol{w} = \boldsymbol{w} + \boldsymbol{v}$ (commutativity)
4. $\boldsymbol{v} + (-1) \cdot \boldsymbol{v} = 0$ (inverse elements)
5. $1 \cdot \boldsymbol{v} = \boldsymbol{v}$ (neutral element of $\cdot$)
6. $s \cdot (t \cdot \boldsymbol{v}) = (s \cdot t) \cdot \boldsymbol{v}$ (compatibility)
7. $s \cdot (\boldsymbol{v} + \boldsymbol{w}) = s \cdot \boldsymbol{v} + s \cdot \boldsymbol{w}$ (distributivity)
8. $(s + t) \cdot \boldsymbol{v} = s \cdot \boldsymbol{v} + t \cdot \boldsymbol{v}$ (distributivity)

Remarks

- The definition is about **real-valued** vector spaces, because we directly use the real numbers $\mathbb{R}$ for the operation $\cdot$.

Remarks

- The definition is about **real-valued** vector spaces, because we directly use the real numbers $\mathbb{R}$ for the operation $\cdot$.
- In mathematics, vector spaces are defined (more generally) over arbitrary fields (e.g. complex numbers).

Remarks

- The definition is about **real-valued** vector spaces, because we directly use the real numbers $\mathbb{R}$ for the operation $\cdot$.
- In mathematics, vector spaces are defined (more generally) over arbitrary fields (e.g. complex numbers).
- In this lecture, we focus thus on a less general abstraction.

Remarks

- The definition is about **real-valued** vector spaces, because we directly use the real numbers $\mathbb{R}$ for the operation $\cdot$.
- In mathematics, vector spaces are defined (more generally) over arbitrary fields (e.g. complex numbers).
- In this lecture, we focus thus on a less general abstraction.
- The elements of $\mathbb{R}$ are called **scalars**, the elements of V **vectors**.

Example 0 – $\mathbb{R}$

The real numbers $\mathbb{R}$ form a real-valued vector space.

Example 0 – $\mathbb{R}$

The real numbers $\mathbb{R}$ form a real-valued vector space.

1. $0 \in \mathbb{R}$ is the zero-vector (neutral element of $+$)
2. $u + (v + w) = (u + v) + w$
3. $v + w = w + v$
4. $v + (-1)v = 0$
5. $1v = v$
6. $s(tv) = (st)v$
7. $s(v + w) = sv + sw$
8. $(s + t)v = sv + tv$

Example 1 – $\mathbb{R}^n$

The most commonly know and most frequently used vector space is $(\mathbb{R}^n, +, \cdot)$ with $n \in \mathbb{N}$ and componentwise $+$ and $\cdot$.

- elements of $\mathbb{R}^n$ have n entries: $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$
- addition: $\boldsymbol{x} + \boldsymbol{y} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}$
- scalar multiplication: $s \cdot \boldsymbol{x} = \begin{pmatrix} s \cdot x_1 \\ s \cdot x_2 \\ \vdots \\ s \cdot x_n \end{pmatrix}$
- $0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

Application Examples

Application Examples

- In a dataset on persons with features size, weight, income, number of children, data instances can be interpreted as vectors in $\mathbb{R}^4$

Application Examples

- In a dataset on persons with features size, weight, income, number of children, data instances can be interpreted as vectors in $\mathbb{R}^4$
- plants with breadth and width of their leafs and petals can be represented as vectors with 4 entries (e.g. IRIS dataset)

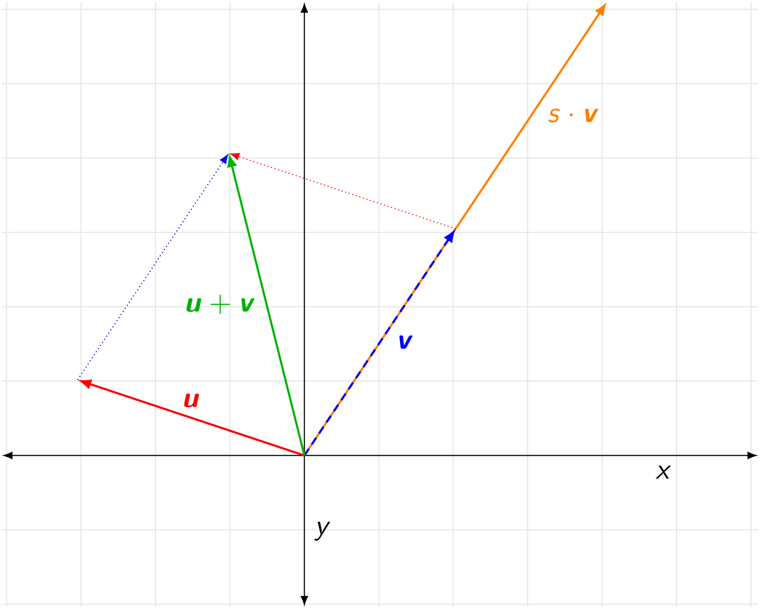
Application Examples

- In a dataset on persons with features size, weight, income, number of children, data instances can be interpreted as vectors in $\mathbb{R}^4$
- plants with breadth and width of their leafs and petals can be represented as vectors with 4 entries (e.g. IRIS dataset)
- the measurements of sensors in a machine at each point in time is a vector in $\mathbb{R}^n$ where n is the number of sensors

Application Examples

- In a dataset on persons with features size, weight, income, number of children, data instances can be interpreted as vectors in $\mathbb{R}^4$
- plants with breadth and width of their leafs and petals can be represented as vectors with 4 entries (e.g. IRIS dataset)
- the measurements of sensors in a machine at each point in time is a vector in $\mathbb{R}^n$ where n is the number of sensors
- bag of words: in text mining: represent each text by a vector of word frequencies; each entry represents one word (dimensionality = number of words in corpus)

Example 1 – $\mathbb{R}^2$ – Geometric Interpretation



Example 1 – $\mathbb{R}^2$ – Geometric Interpretation

- The geometric interpretation works similarly in $\mathbb{R}^n$ with $n > 2$.
- It allows computing various geometric entities, like planes, volumes, distances, angles, ...

Notebook 04_1_vectors_in_python, Cells 1–6

Example 1 – $\mathbb{R}^n$: Vectors vs. Tuples

- A tuple is a finite, ordered list of elements.
- Vectors in $\mathbb{R}^n$ are n-dimensional tuples.

Example 1 – $\mathbb{R}^n$: Vectors vs. Tuples

- A tuple is a finite, ordered list of elements.
- Vectors in $\mathbb{R}^n$ are n-dimensional tuples.
- However: Not all tuples are vectors, not all vectors are tuples!

Example 1 – $\mathbb{R}^n$: Vectors vs. Tuples

- A tuple is a finite, ordered list of elements.
- Vectors in $\mathbb{R}^n$ are n-dimensional tuples.
- However: Not all tuples are vectors, not all vectors are tuples!
- E.g. polynomials form a vector space (infinite dimensions), but its elements are no tuples.

Example 1 – $\mathbb{R}^n$: Vectors vs. Tuples

- A tuple is a finite, ordered list of elements.
- Vectors in $\mathbb{R}^n$ are n-dimensional tuples.
- However: Not all tuples are vectors, not all vectors are tuples!
- E.g. polynomials form a vector space (infinite dimensions), but its elements are no tuples.
- E.g. tuples with categorical attributes, tuples with integer attributes (can be interpreted as vectors, but form no vector space)

Example 2 – $\mathbb{R}^{n \times m}$

Definition 2 (Matrix)

Let $m, n \in \mathbb{N}$. An $m \times n$ dimensional matrix M is a tuple with $m \cdot n$ elements m_{ij} $(1 \leq i \leq m,\ 1 \leq j \leq n)$.

Matrixes are usually displayed as two-dimensional arrays:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

- addition and multiplication elementwise

Notebook 04_1_vectors_in_python, Cells 7–11

Exercises 1–3

Handling Tabular Data

- Pandas library

Notebook 04_2_tabular_data