Technische Universität Berlin

Mathematical foundations for machine learning

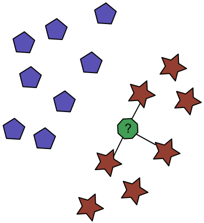
July 29, 2019

Machine Learning - Learning from data

Approach:

1. Extraction of features and representation as vectors
2. Application of methods from linear algebra, stochastics, ..., to learn from data
3. Application of learned model

Example: k-nearest-neighbours-classification



- training data and corresponding labels are given
- a previously unknown data point (test data) ϑ receives the label of its k nearest neighbors

Exercise 1

A set of training data is given (red and blue). Using kNN to classify the previously unknown data point (green). Which label will be assigned for $k \in 1, 2, 5$?

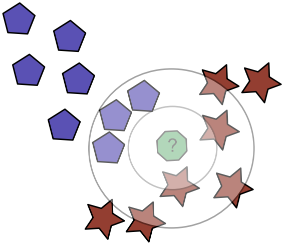


Table of Contents

Vector Spaces

Definition. Let K be a field. A set V together with two operations:

vector addition

$$\dot{+} : V \times V \to V \qquad (\vartheta, \omega) \mapsto \vartheta \dot{+} \omega$$

and scalar multiplication

$$\cdot : K \times V \to V \qquad (\lambda, \vartheta) \mapsto \lambda \cdot \vartheta$$

is called vector space, if

- (V,$\dot{+}$) forms an abelian group with an identity element called zero vector $\mathbf{0}$.
- $\forall \vartheta, \omega \in V$ and $\forall \lambda, \mu \in K$ it is the case that:
  - **i)** $\lambda(\mu\vartheta) = (\lambda\mu)\vartheta$ (associative)
  - **ii)** $1\vartheta = \vartheta$ (identity element)
  - **iii)** $\lambda(\vartheta \dot{+} \omega) = \lambda\vartheta \dot{+} \lambda\omega$, $(\lambda + \mu)\vartheta = \lambda\vartheta \dot{+} \mu\vartheta$ (distributive)

Linear Combination

Definition. Let V be a $K-$vector space, $\vartheta_1, \ldots, \vartheta_n \in V$, $n \in \mathbb{N}$ and $\lambda_1, \ldots, \lambda_n \in K$.
Then $\forall \vartheta \in V$ that can be represented as

$$\vartheta = \lambda_1 \vartheta_1 + \ldots + \lambda_n \vartheta_n,$$

ϑ is defined as a linear combination of $\vartheta_1, \ldots, \vartheta_n$.

Linear Independence

Definition. Let V be a $K-$vector space. A set of vectors $\vartheta_1, \ldots, \vartheta_n \in V$ is called linearly independent, if from

$$\lambda_1, \ldots, \lambda_n \in K \text{ and } \lambda_1\vartheta_1 + \ldots + \lambda_n\vartheta_n = 0$$

it follows that

$$\lambda_1 = \ldots = \lambda_n = 0.$$

If no vector in the set is a linear combination of the other vectors.

Table of Contents

Inner Product

Definition. Let V be a $\mathbb{R}-$vector space. An inner product on V defines a mapping $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ with the following properties:

i) bilinear: $\langle \lambda\vartheta, \omega \rangle = \lambda \langle \vartheta, \omega \rangle = \langle \vartheta, \lambda\omega \rangle$,
$\langle \vartheta + \upsilon, \omega \rangle = \langle \vartheta, \omega \rangle + \langle \upsilon, \omega \rangle$, $\langle \vartheta, \omega + \upsilon \rangle = \langle \vartheta, \omega \rangle + \langle \vartheta, \upsilon \rangle$

ii) symmetric: $\langle \vartheta, \omega \rangle = \langle \omega, \vartheta \rangle$

iii) positive definite: $\langle \vartheta, \vartheta \rangle \geq 0$ and $\langle \vartheta, \vartheta \rangle = 0 \Leftrightarrow \vartheta = \mathbf{0}$

$\forall \upsilon, \vartheta, \omega \in V$ and $\lambda \in \mathbb{R}$

Inner Product: Example

- dot product on $\mathbb{R}^n$:

$$\langle \mathbf{x}, \mathbf{y} \rangle := x_1 y_1 + \ldots + x_n y_n = \mathbf{x}^\top \mathbf{y} \qquad (\mathbf{x}, \mathbf{y} \in \mathbb{R}^n)$$

- inner product on $\mathbb{R}^n$ with $\lambda_1, \ldots, \lambda_n \geq 0$:

$$\langle \mathbf{x}, \mathbf{y} \rangle := \lambda_1 x_1 y_1 + \ldots + \lambda_n x_n y_n \qquad (\mathbf{x}, \mathbf{y} \in \mathbb{R}^n)$$

- inner product on $\mathcal{F} = L_2(X) = \{f : X \to \mathbb{R} \mid \int_X f(x)^2 dx < \infty\}$, the space of quadratic integrable functions on a compact $X \subset \mathbb{R}^n$

$$\langle f, g \rangle := \int_X f(x) g(x) dx \qquad (f, g \in \mathcal{F})$$

Norm - Length of a vector

Definition. Let V be a $\mathbb{R}-$vector space. A norm is a mapping $\|\cdot\|: V \to \mathbb{R}^+$ with the following properties

i) $\|\vartheta\| \geq 0$ and $\|\vartheta\| = 0 \Leftrightarrow \vartheta = \mathbf{0}$

ii) $\|\vartheta + \omega\| \leq \|\vartheta\| + \|\omega\|$ (triangle inequality)

iii) $\|\lambda\vartheta\| = |\lambda|\|\vartheta\|$

$\forall \vartheta, \omega \in V, \; \lambda \in \mathbb{R}$

examples on $\mathbb{R}^n$:

- 2-norm $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$
- 1-norm $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$
- max-norm $\|\mathbf{x}\|_\infty = \max_i |x_i|$

Exercise 2

1. We can generalize these norms to the p-norm
   $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$
   For p = 1 we get the taxicab norm, for p = 2 we get the Euclidean norm.
   Show that as p approaches ∞ the p-norm approaches the infinity norm or maximum norm.

   $$\lim_{p\to\infty} \|\mathbf{x}\|_p = \max_i |x_i|$$

   Hint: take lower and upper bounds.

Exercise 3

1. In the xy-plane sketch all points for which the 2-Norm is 1.

$$C_2 := \{(x, y)^T \in \mathbb{R}^2 | \, \|(x, y)^T\|_2 = x^2 + y^2 = 1\}.$$

2. Sketch all points for which the 1-Norm is 1.

$$C_1 := \{(x, y)^T \in \mathbb{R}^2 | \, \|(x, y)^T\|_1 = |x| + |y| = 1\}.$$

3. Sketch all points for which the ∞-Norm is 1.

$$C_\infty := \{(x, y)^T \in \mathbb{R}^2 | \, \|(x, y)^T\|_\infty = max(|x|, |y|) = 1\}.$$

Euclidean Vector Spaces

Definition. A $\mathbb{R}-$vector space together with an inner product $(V, \langle \cdot, \cdot \rangle)$ is defined as an euclidean vector space.

In an euclidean vector space the norm is introduced by the euclidean norm:

$$\| \cdot \| : V \to \mathbb{R}^+ \qquad \|v\| = \sqrt{\langle v, v \rangle}$$

Orthogonal Vectors

Definition. Two vectors ϑ, ω of an euclidean vector space are called orthogonal to each other ($\vartheta \perp \omega$) if $\langle \vartheta, \omega \rangle = 0$.

Definition. A set of vectors $\vartheta_1, \ldots, \vartheta_n$ of an euclidean vector space builds an orthonormal set if $\|\vartheta_i\| = 1 \forall i$ and $\vartheta_i \perp \vartheta_j$ for $i \neq j$.

Table of Contents

Matrices

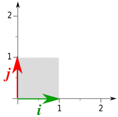
Definition. A real $m \times n$ matrix is an array of $m \cdot n$ elements of $\mathbb{R}$ with the following pattern:

$$\begin{pmatrix} a_{1,1} & \cdots\cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots\cdots & a_{m,n} \end{pmatrix}$$

where $a_{i,j}$ are the coefficients of the matrix.

Exercise 4

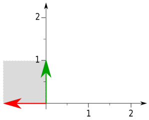
Imagine the $\mathbb{R}^2$ space with unit vectors
$i = (1, 0)^T$ and $j = (0, 1)^T$

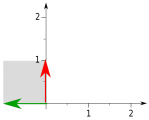


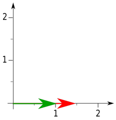
- What do these matrices do to space:

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1.5 & 0 \\ 0 & 2 \end{pmatrix}$$

- What matrices correspond to these transformations







Determinant

Theorem and Definition. There exists exactly one mapping $\det : \mathbb{R}^{n\times n} \to \mathbb{R},\ A \mapsto \det(A)$ with

i) det is linear in each row (in each column)

ii) If the (column-) rank is smaller than n, the $\det(A) = 0$

iii) $\det(I_n) = 1$

This mapping is called determinant

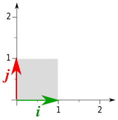
Determinant

The determinant of a quadratic matrix is a number. What does it mean?

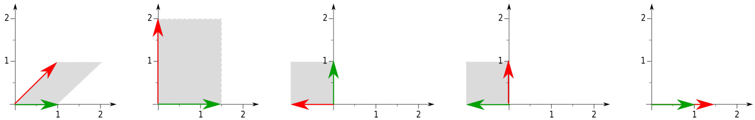
- absolute value: the volume of a parallelotope spannend by the row/column vectors of the matrix
- sign: orientation of the parallelotope

Exercise 5

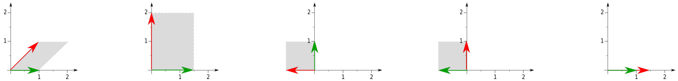
Imagine the $\mathbb{R}^2$ space with unit vectors $i = (1,0)^T$ and $j = (0,1)^T$



What are the deterinantes of the matrice that transform space in the following ways?



Matrix Examples



matrix:	$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1.5 & 0 \\ 0 & 2 \end{pmatrix}$	$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1.5 \\ 0 & 0 \end{pmatrix}$
det.:	1	3	1	-1	0

Determinant

Let $A, B \in \mathbb{R}^{n\times n}$ be two quadratic matrices and $\lambda \in \mathbb{R}$

- If A is a triangle matrix, then the determinant is the product of the elements along the main diagonal of A.

- A is invertible if and only if $\det A \neq 0$.

- $\det AB = \det A \det B$
- $\det A^{-1} = (\det A)^{-1}$
- $\det A = \det A^T$
- $\det \lambda A = \lambda^n \det A$

- $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$

Rank

Definition. The column/row rank of a matrix is the maximum number of linearly independent columns/rows.

Theorem. column rank $=$ row rank

Matrix Inverse

Definition. A square matrix $A \in \mathbb{R}^{n\times n}$ is invertible (regular, non-singular), if there is another matrix A^{-1}, such that

$$AA^{-1} = A^{-1}A = I_n.$$

Useful properties:

- A is invertible if and only if $\operatorname{rank}(A) = n$
- $(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$
- $(A^T)^{-1} = (A^{-1})^T$
- $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

Exercise 6

3. For every square $n \times n$ Matrix A:

- $\operatorname{rank}(A) = n \Rightarrow A$ is invertible, but there are invertible A with $\operatorname{rank}(A) \neq n$
- A is invertible $\Rightarrow \operatorname{rank}(A) = n$, but there is an A with $\operatorname{rank} A = n$, which is not invertible.
- $\operatorname{rank} A = n \Leftrightarrow A$ is invertible

4. The rank of the matrix $\begin{pmatrix} 4 & 4 & 4 \\ 4 & 4 & 4 \\ 4 & 4 & 4 \end{pmatrix}$ is

- 1
- 3
- 4

Trace

Definition and Theorem. The trace of a quadratic matrix $C \in \mathbb{R}^{n \times n}$ is the sum of its elements in the main diagonal.

$$\mathrm{Tr}(C) = \sum_{i}^{n} c_{ii}$$

for $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$$

Matrix Transpose

Definition. Let $A \in \mathbb{R}^{m \times n}$ be a matrix. The matrix transpose of A is defined as:

$$A^\top = (a_{ij})^\top = a_{ji} \quad \text{for} \begin{cases} i = 1, \ldots, m \\ j = 1, \ldots, n. \end{cases}$$

It holds that:

i) $(A + B)^\top = A^\top + B^\top$

ii) $(\lambda \cdot A)^\top = \lambda \cdot A^\top$

iii) $(A \cdot C)^\top = C^\top \cdot A^\top$

Special Matrices

Let $A \in \mathbb{R}^{n \times n}$ be a quadratic, real matrix.

- A is **orthogonal** if $AA^T = A^T A = I_n$.
  Orthogonal matrices represent reflections and rotations in space.
- A is **symmetric** if $A = A^T$.
- A is **anti symmetric** if $A = -A^T$.
- A is **diagonal** if all elements except for those in the main diagonal are 0.

Exercise 7

Let $R \in \mathbb{R}^{n \times n}$ be an orthogonal matrix that is to say $RR^T = R^T R = I$. Show that the dot product of two vectors is invariant to a multiplication with R which means that for all $\mathbf{x}, \mathbf{y} \in R^n$ it is the case that

$$\langle R\mathbf{x}, R\mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle .$$

Table of Contents

Eigenvalues & Eigenvectors

Definition. An eigenvector of a square matrix $A \in \mathbb{R}^{n \times n}$ to an eigenvalue $\lambda \in \mathbb{C}$ is a vector $\mathbf{v} \in \mathbb{R}^n \setminus \{0\}$ such that

$$A\mathbf{v} = \lambda \mathbf{v}.$$

Calculation

Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. Then for every eigenvector $\mathbf{v}$ it holds that

$$A\mathbf{v} = \lambda \mathbf{v} \Leftrightarrow (A - \lambda I)\mathbf{v} = 0$$

$\Rightarrow (A - \lambda I)$ has to be singular

1. Calculate the eigenvalues as the roots of the characteristic polynomial $P(\lambda) := \det(A - \lambda I)$.
2. Find for every real eigenvalue λ_i a basis of the vector space $\{\mathbf{v} \in \mathbb{R}^n | (A - \lambda_i I)\mathbf{v} = 0\}$.

Let $A \in \mathbb{R}^{n \times n}$ be a square matrix:

- there is at most n real eigenvalues and at most n linearly independent eigenvectors.
- there can be less than n linearly independent eigenvectors, even if there are n real roots of the characteristic polynomial.
- pairwise eigenvectors which differ in their corresponding eigenvalues are linearly independent.
- A has n differing eigenvalues $\Rightarrow$ A has n linearly independent eigenvectors.

Exercise 8

Let $A \in \mathbb{R}^{n \times n}$ be a square matrix and $p \in \mathbb{N}$. Then A and A^p will have

- the same eigenvectors and eigenvalues
- the same eigenvalues, but not necessarily the same eigenvectors
- the same eigenvectors, but not necessarily the same eigenvalues.

Table of Contents

Diagonalizable Matrices

Definition. A square matrix $A \in \mathbb{R}^{n\times n}$ is called diagonalizable, if there exists an invertible matrix $S \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$, such that

$$\Lambda = S^{-1}AS.$$

$A = S\Lambda S^{-1}$ is also called eigendecomposition of A.

considerable properties:

- A has n linearly independent eigenvectors $\Leftrightarrow$ A diagonalizable
- the columns of S are the eigenvectors of A, the diagonal of Λ contains the corresponding eigenvalues.

Properties of diagonalizable matrices

Let $A \in \mathbb{R}^{n\times n}$ be diagonalizable with $A = S\Lambda S^{-1}$.

- all eigenvalues $\neq 0 \Leftrightarrow A$ is invertible and

$$A^{-1} = (S\Lambda S^{-1})^{-1} = S\Lambda^{-1}S^{-1}$$

- simplify diagonalization by exponentiation of A:

$$A^p = S\Lambda^p S^{-1} \text{ für } p \in \mathbb{N}$$

- the determinant is the product of the eigenvalues.
- the trace is the sum of the eigenvalues.

Table of Contents

Symmetric Matrices

For every **symmetric** matrix $A \in \mathbb{R}^{n \times n}$ it is the case:

- its eigenvalues are real.
- there are always n orthogonal eigenvectors.
- A can be decomposed into

$$A = U \Lambda U^T$$

where U is an orthogonal matrix with eigenvectors as columns and Λ is the diagonal matrix with the corresponding eigenvalues on its diagonal.

Positive Definiteness

Definition. a square matrix $A \in \mathbb{R}^{n\times n}$ is called

positive definite	if $\mathbf{v}^T A\mathbf{v} > 0$
positive semidefinite	if $\mathbf{v}^T A\mathbf{v} \geq 0$
negative definite	if $\mathbf{v}^T A\mathbf{v} < 0$
negative semidefinite	if $\mathbf{v}^T A\mathbf{v} \leq 0$
	$\forall \mathbf{v} \in \mathbb{R}^n \setminus \{0\}$

Theorem. For every symmetric matrix A:

A positive definite	$\Leftrightarrow$	all eigenvalues > 0
A positive semidefinite	$\Leftrightarrow$	all eigenvalues ≥ 0
A negative definite	$\Leftrightarrow$	all eigenvalues < 0
A negative semidefinit e	$\Leftrightarrow$	all eigenvalues ≤ 0

- Many important equations for calculating matrices: K. B. Petersen, M. S. Pedersen (2007) The Matrix Cookbook. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274

Table of Contents

Definition. $f : \mathbb{R} \supset J \to \mathbb{R}$ is called differentiable in $x_0 \in J$, if

$$f'(x) := \lim_{y \to x} \frac{f(y) - f(x)}{y - x} \quad \text{exists in } \mathbb{R}$$

$f'(x)$ is called derivative of f in x.
f is differentiable on J, if it is differentiable for all $x \in J$.
alternatively: $h := y - x$

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Example. $n \in \mathbb{N}, \quad f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto x^2$

$$
\begin{aligned}
f'(x) &= \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} \\
&= \lim_{h \to 0} \frac{(x^2 + 2xh + h^2) - x^2}{h} \\
&= \lim_{h \to 0} \frac{2xh + h^2}{h} \\
&= \lim_{h \to 0} (2x + h) \\
&= 2x
\end{aligned}
$$

Geometric Interpretation.

$$\frac{f(x)-f(x_0)}{x-x_0} \qquad \text{slope of secant}$$

$$\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0} \qquad \text{slope of tangent}$$

Rules of Differentiation I

Theorem. Let $f, g : V \to \mathbb{R}$ in $x \in V$ be differentiable functions and $\lambda \in \mathbb{R}$. Then the functions $f + g, \lambda f, f \cdot g$ in x are differentiable and the following rules apply:

i) linearity

$$(f+g)'(x) = f'(x) + g'(x)$$
$$(\lambda f)'(x) = \lambda f'(x)$$

ii) product rule

$$(f \cdot g)'(x) = f'(x)g(x) + f(x)g'(x)$$

(we can also derive the rule for integration by parts from this:)

$$[f(x) \cdot g(x)]_a^b = \int_a^b f'(x)g(x)dx + \int_a^b f(x)g'(x)dx$$

Rules of Differentiation II

Theorem. If it is also the case that $g(\varphi) \neq 0 \ \forall \varphi \in V$, it follows that $(f/g) : V \to \mathbb{R}$ in x is differentiable and the following rule applies:

iii) quotient rule

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}.$$

or just apply product rule (and chain rule)

$$\begin{aligned}\left(f(x)g(x)^{-1}\right)' &= f'(x)g(x)^{-1} + f(x)(-1)g(x)^{-2}g'(x) \\ &= \frac{f'(x)}{g(x)} - \frac{f(x)g'(x)}{g(x)^2}\end{aligned}$$

Rules of Differentiation III

Theorem. Let $f : V \to \mathbb{R}$ and $g : W \to \mathbb{R}$ be functions with $f(V) \subset W$, where f is differentiable in $x \in V$ and g is differentiable in $y := f(x) \in W$. Then the composite function $g \circ f : V \to \mathbb{R}$ is differentiable in x and the following rule applies:

iv) chain rule

$$(g \circ f)'(x) = g'(f(x))f'(x).$$

Exercise 9

We consider two functions f and g which transform an input vector $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$ into a scalar:
$f(\mathbf{x}) = \mathbf{u}^\top \mathbf{x},\ \mathbf{u} = (u_1, \ldots, u_d)^\top \in \mathbb{R}^d$ and $g(\mathbf{x}) = \mathbf{x}^\top \mathbf{x}$.

- Compute the partial derivative of f with respect to one entry x_j ($j \in \{1, 2, \ldots, d\}$), that is $\frac{\partial f(\mathbf{x})}{\partial x_j}$
- Compute the gradient $\nabla g(\mathbf{x}) = \left(\frac{\partial g(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial g(\mathbf{x})}{\partial x_d}\right)^\top$ for g.