INTELLIGENTE DATENANALYSE IN MATLAB

Mathematische Grundlagen

Michael Brückner/Tobias Scheffer

Literatur

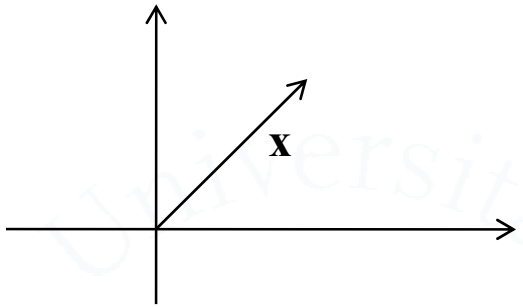
- A. Fischer, K. Vetters: Lineare Algebra – Eine Einführung für Ingenieure und Naturwissenschaftler.
- H. Amann, J. Escher: Analysis I-III.
- S. Boyd, L. Vandenbergh: Convex Optimization.
- R. Schlittgen: Einführung in die Statistik.
- H. R. Schwarz: Numerische Mathematik.

Überblick

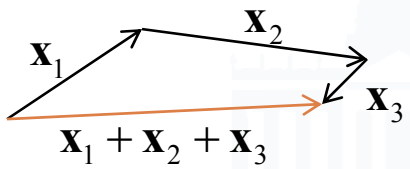
- Lineare Algebra
- Analysis
- Stochastik
- Numerik

Lineare Algebra: Vektoren

- Vektor: $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix}^{\mathrm{T}}$

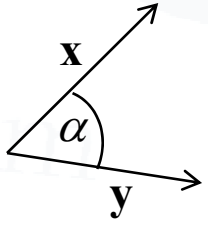


- Vektorsumme: $\sum_{i=1}^{n} \mathbf{x}_i = \begin{bmatrix} x_{11} + \ldots + x_{n1} \\ \vdots \\ x_{1m} + \ldots + x_{nm} \end{bmatrix}$



- Skalarprodukt: $\langle \mathbf{y}, \mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\mathrm{T}} \mathbf{y} = \sum_{i=1}^{m} x_i y_i$

$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \|\mathbf{y}\| \cos \alpha$



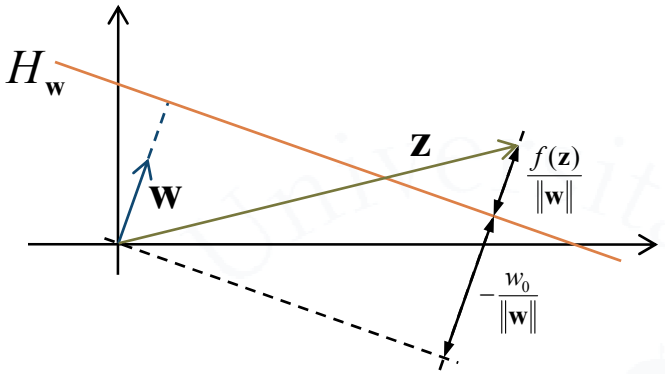
Lineare Algebra: Matrizen

- Matrix: $$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{m1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{mn} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}$$
- Matrixsumme: $$\mathbf{X} + \mathbf{Y} = \begin{bmatrix} x_{11} + y_{11} & \cdots & x_{1n} + y_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} + y_{m1} & \cdots & x_{mn} + y_{mn} \end{bmatrix}$$
- Matrixprodukt: $$\mathbf{YX} \neq \mathbf{XY} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nk} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_{1i} y_{i1} & \cdots & \sum_{i=1}^{n} x_{1i} y_{ik} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{mi} y_{i1} & \cdots & \sum_{i=1}^{n} x_{mi} y_{ik} \end{bmatrix}$$

Lineare Algebra: Geometrie

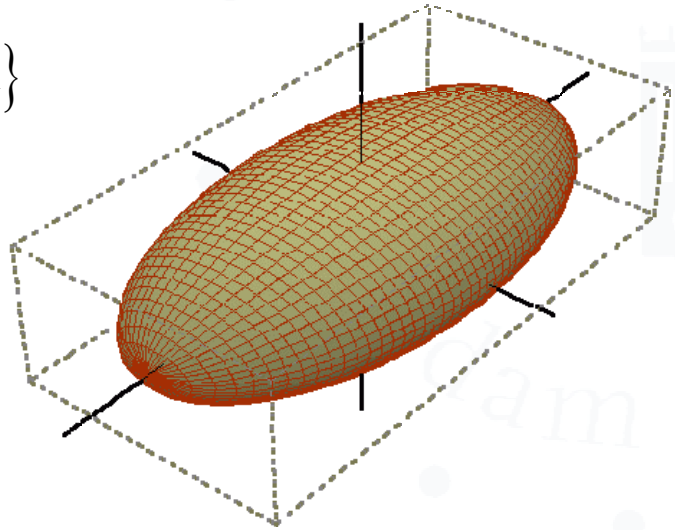
- Hyperebene:

$$H_{\mathbf{w}} = \left\{ \mathbf{x} \mid f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{w} + w_0 = 0 \right\}$$



- Ellipsoid:

$$E_{\mathbf{A}} = \left\{ \mathbf{x} \mid g(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = 1 \right\}$$



Lineare Algebra: Matrix-Eigenschaften

- Quadratisch: $n = m$
- Symmetrisch: $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$
- Spur (trace): $tr(\mathbf{A}) = \sum_{i=1}^{m} a_{ii}$
- Rang (rank): $rk(\mathbf{A}) =$ Anzahl linear unabhänger Zielen/Spalten
- Determinante: $det(\mathbf{A}) = vol(E_{\mathbf{A}})^{-2}$
- Positiv definit: $\forall \mathbf{x} \neq \mathbf{0} : \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x} > 0$

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

gilt nur falls $\mathbf{A}$ positiv definit

äquivalent gilt $\exists \mathbf{G} : \mathbf{A} = \mathbf{G}\mathbf{G}^{\mathrm{T}}$

Lineare Algebra: Spezielle Matrizen

- Eins-Vektor/-Matrix: $\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$
- Einheitsvektor: $\mathbf{e}_i = [\underbrace{0 \quad \cdots \quad 0}_{i-1} \quad 1 \quad 0 \quad \cdots \quad 0]^T$
- Diagonalmatrix: $diag(\mathbf{a}) = [a_1\mathbf{e}_1 \quad \cdots \quad a_m\mathbf{e}_m] = \begin{bmatrix} a_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_m \end{bmatrix}$
- Einheitsmatrix: $\mathbf{I} = diag(\mathbf{1}) = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$

Lineare Algebra: Matrix-Faktorisierung

- LU-Zerlegung $(m = n)$: $\mathbf{A} = \mathbf{LU} = \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ l_{m1} & \cdots & l_{mm} \end{bmatrix} \begin{bmatrix} u_{11} & \cdots & u_{m1} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & u_{mm} \end{bmatrix}^{\mathrm{T}}$
- Cholesky-Zerlegung $(m = n)$: $\mathbf{A} = \mathbf{G}\mathbf{G}^{\mathrm{T}}$

existiert nur falls $\mathbf{A}$ positiv definit

- Eigenwert-Zerlegung $(m = n)$:

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_m \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_m \end{bmatrix}^{\mathrm{T}} \quad \mathbf{v}_i^{\mathrm{T}} \mathbf{v}_j = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j \end{cases}$$

Eigenvektoren

Eigenwerte

Lineare Algebra: Matrix-Faktorisierung

- Singulärwert-Zerlegung ($m > n$):

Singulärwerte

$$\mathbf{A} = \mathbf{U}\mathbf{\Omega}\mathbf{V}^{\mathrm{T}} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_m\end{bmatrix} \left[\begin{array}{ccc}\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \\ \hline \multicolumn{3}{c}{\mathbf{0}}\end{array}\right] \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_n\end{bmatrix}^{\mathrm{T}}$$

$$\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j = \begin{cases}1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j\end{cases}$$

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = \begin{cases}1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j\end{cases}$$

- Berechnung durch Eigenwert-Zerlegung von $\mathbf{A}^{\mathrm{T}}\mathbf{A}$:

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{V}\begin{bmatrix}\lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n\end{bmatrix}\mathbf{V}^{\mathrm{T}}, \quad \mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{U}\left[\begin{array}{ccc|c}\lambda_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & \cdots & \lambda_n & \\ \hline \multicolumn{3}{c|}{\mathbf{0}} & \mathbf{0}\end{array}\right]\mathbf{U}^{\mathrm{T}}, \quad \sigma_i = \sqrt{\lambda_i}$$

Analysis: Distanzen

- Definition: $d(x,y)=0 \Leftrightarrow x=y \quad d(x,y)=d(y,x) \quad d(x,y)\le d(x,z)+d(z,y)$
- Beispiele für Vektor-Distanzen
  - Minkowski-Distanz: $\|\mathbf{x}-\mathbf{y}\|_p = \sqrt[p]{\sum_{i=1}^{m}|x_i-y_i|^p}$
  - Manhattan-Distanz: $\|\mathbf{x}-\mathbf{y}\|_1$
  - Euklidische Distanz: $\|\mathbf{x}-\mathbf{y}\|_2$
- Beispiel für Matrix-Distanzen:
  - Schatten-Distanz: $\|\mathbf{X}-\mathbf{Y}\|_p = \sqrt[p]{\sum_{i=1}^{m}\sigma_i^p}$
  - Trace-Distanz: $\|\mathbf{X}-\mathbf{Y}\|_{tr} = \|\mathbf{X}-\mathbf{Y}\|_1$
  - Frobenius-Distanz: $\|\mathbf{X}-\mathbf{Y}\|_F = \|\mathbf{X}-\mathbf{Y}\|_2$

Norm von x: $\|x\| = d(x,0)$

Singulärwerte der Matrix $\mathbf{X}-\mathbf{Y}$

Analysis: Differentialrechnung

- Erste Ableitung einer Funktion f:
  - Nach einem Skalar x: $\nabla_x f = \frac{\partial f}{\partial x}$
  - Nach einem Vektor $\mathbf{x}$: $\nabla_{\mathbf{x}} f = grad(f) = \left[\frac{\partial f}{\partial x_1} \quad \cdots \quad \frac{\partial f}{\partial x_m} \right]^T$

Gradient

Partielle Ableitung

- Zweite Ableitung einer Funktion f:
  - Nach einem Skalar x: $\nabla_x^2 f = \frac{\partial^2 f}{\partial x^2}$
  - Nach einem Vektor $\mathbf{x}$: $\nabla_{\mathbf{x}}^2 f = H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix}$

Hesse-Matrix

Analysis: Integralrechnung

- Integral einer Funktion f:
  - Über einem Skalar x: $F_x = \int f(x) \partial x$
  - Über einem Vektor $\mathbf{x}$: $F_{\mathbf{x}} = \int f(\mathbf{x}) \partial \mathbf{x} = \int \cdots \int f(\mathbf{x}) \partial x_1 \cdots \partial x_m$
- Bestimmtes Integral: $\int_a^b f(x) \partial x = F_x(b) - F_x(a)$
- Umkehroperation: $f(x) = \frac{\partial F_x}{\partial x}$
- Berechnung analytisch durch Integrationsregeln oder numerische Approximation (Quadraturformeln).

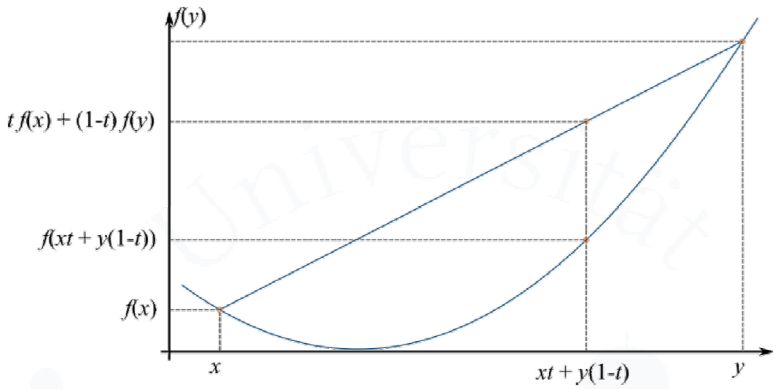
Analysis: Konvexität

- Konvexe Funktion f:

$$f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)$$

- Konkave Funktion f:

$$f(tx+(1-t)y) \geq tf(x)+(1-t)f(y)$$



- Streng konvex bzw. konkav:
  - „≤“ bzw. „≥“ wird zu „<“ bzw. „>“.
  - Es existiert genau ein Minimum bzw. Maximum.
  - Zweite Ableitung ist überall positiv bzw. negativ.
  - Tangente an $f(x)$ ist untere bzw. obere Schranke von f.

Analysis: Optimierung

- Optimierungsaufgabe (OA): $f^* = \min_{x \in S} f(x) \;\text{ mit }\; x^* = \arg\min_{x \in S} f(x)$
  - f Zielfunktion.
  - S zulässiger Bereich (definiert durch Nebenbedingungen).
  - f^* Optimalwert.
  - x^* optimale Lösung.
  - Ein $x \in S$ wird *zulässige Lösung* genannt.
- Konvexe OA:
  - Zielfunktion und zulässiger Bereich konvex.
  - Lokales Optimum = Globales Optimum.

Analysis: Optimierung

- Notwendige Optimalitätskriterien für x^*:
  - Wenn f in x^* differenzierbar ist, dann ist $\nabla_x f(x^*) = 0$.
  - Wenn f in x^* zweimal differenzierbar ist, dann ist $\nabla_x^2 f(x^*)$ eine positiv (semi-)definite Matrix.
- OA ohne Nebenbedingungen:

$$S = \mathbb{R}^m$$

- OA mit n Nebenbedingungen:

$$S = \left\{ \mathbf{x} \in \mathbb{R}^m \mid g_i(\mathbf{x}) \leq 0,\ g_j(\mathbf{x}) = 0,\ i = 1 \ldots k,\ j = k+1 \ldots n \right\}$$

Analysis: Optimierung

- Lagrange-Ansatz für OA mit Nebenbedingungen:
  - Nebenbed.: $S = \left\{ \mathbf{x} \in \mathbb{R}^m \mid g_i(\mathbf{x}) \leq 0,\ g_j(\mathbf{x}) = 0,\ i = 1 \ldots k,\ j = k+1 \ldots n \right\}$
  - Lagrange-Funktion: $L(\mathbf{x}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i g_i(\mathbf{x})$
  - Dualität: $f^* = \min_{\mathbf{x} \in S} f(\mathbf{x}) = \underbrace{\min_{\mathbf{x} \in \mathbb{R}^m} \max_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\alpha})}_{f_p(\mathbf{x})} \geq \underbrace{\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \min_{\mathbf{x} \in \mathbb{R}^m} L(\mathbf{x}, \boldsymbol{\alpha})}_{f_d(\boldsymbol{\alpha})}$

    Dualitätslücke
  - Primale OA: $\min_{x \in \mathbb{R}^m} f_p(\mathbf{x}) \quad \text{mit} f_p(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{falls } \mathbf{x} \in S \\ \infty & \text{falls } \mathbf{x} \notin S \end{cases}$
  - Duale OA: $\max_{\boldsymbol{\alpha} \geq \mathbf{0}} f_d(\boldsymbol{\alpha}) \quad \text{mit} f_d(\boldsymbol{\alpha}) = \min_{x \in \mathbb{R}^m} L(\mathbf{x}, \boldsymbol{\alpha})$

Stochastik: Wahrscheinlichkeitstheorie

- Zufallsexperiment: definierter Prozess, in dem eine Beobachtung ω erzeugt wird (Elementarereignis).
- Ereignisraum Ω: Menge aller möglichen Elementar-ereignisse.
- Ereignis A: Teilmenge des Ereignisraums.
- Wahrscheinlichkeitsfunktion P: Funktion welche Wahrscheinlichkeitsmasse auf Ereignisse A aus Ω verteilt.

Stochastik: Wahrscheinlichkeitstheorie

- Gültige Wahrscheinlichkeitsfkt. (Kolmogorow-Axiome)
  - Wahrscheinlichkeit von Ereignis $A \subseteq \Omega$: $0 \le P(A) \le 1$
  - Sicheres Ereignis: $P(\Omega) = 1$
  - Für die Wahrscheinlichkeit zweier *unabhängiger* (inkompatibler) Ereignisse $A \subseteq \Omega$ und $B \subseteq \Omega$ (d.h. $A \cap B = \varnothing$) gilt: $P(A \cup B) = P(A) + P(B)$
- Summenregel: $P(A) = \sum_i P(A \cap B_i)$ — $\{B_i\}$ ist Partitionierung von Ω
- Produktregel: $P(A \cap B) = P(A \mid B) P(B)$
- Satz von Bayes: $P(A \mid B) P(B) = P(B \mid A) P(A) \Leftrightarrow P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}$

Stochastik: Wahrscheinlichkeitstheorie

- Zufallsvariable X: Abbildung eines elementaren Ereignisses auf einen numerischen Wert, $X : \omega \in \Omega \mapsto x \in \mathbb{R}$.
  - Elementarereignis $\omega \leftrightarrow$ Belegung der Zufallsvariable $X(\omega)=x$.
- Verteilungsfunktion einer Zufallsvariable X:

$$F_X(x) = P(X \le x) = P(\{\omega \in \Omega \mid X(\omega) \le x\})$$

- Dichtefunktion einer Zufallsvariable X:

$$f_X(x) = P(X = x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

- Zusammenhang von Verteilungs- und Dichtefunktion:

$$F_X(a) = \int_{-\infty}^{a} f_X(x)\partial x \quad \Leftrightarrow \quad f_X(a) = \frac{\partial F_X(a)}{\partial x}$$

Stochastik: Informationstheorie

- Informationsgehalt der Realisierung x einer Zufallsvariable X: $h(x) = I(X = x)$
  - Idee: *Information* zweier unabhängiger Ereignisse soll sich addieren, $h(x, y) = I(X = x) + I(Y = y)$.
  - Für zwei unabhängige Ereignisse gilt

$$p(x, y) = P(X = x \cap Y = y) = P(X = x)P(Y = y)$$

    und somit $h(x, y) = -\log p(x, y)$ mit $h(x) = I(X = x) = -\log P(X = x)$.
  - Für bedingte Ereignisse gilt: $h(x, y) = h(x \mid y) + h(y)$
  - Analog zum Satz von Bayes gilt:

$$h(x \mid y) + h(y) = h(y \mid x) + h(x) \Leftrightarrow h(x \mid y) = h(x, y) - h(y)$$

Stochastik: Kenngrößen von Zufallsvariablen

- Verteilung/Dichte.
- Wertebereich: stetig/diskret, endlich/unendlich, ...
- Erwartungswert (mittlere Realisierung):

$$\mu_X = \mathrm{E}[X] = \sum_x p(x)x$$

- Varianz (mittlere quadratische Abweichung vom Erwartungswert):

$$\sigma_X^2 = \mathrm{E}\left[\left(X - \mu_X\right)^2\right] = \sum_x p(x)\left(x - \mu_X\right)^2$$

- Entropie (mittlerer Informationsgehalt):

$$\mathrm{H}_X = \mathrm{E}[h(X)] = -\sum_x p(x)\log p(x)$$

Stochastik: Mathematische Statistik

- Annahme: Daten (Stichprobe) = Realisierungen bzw. Belegungen von Zufallsvariablen.
- Ziel: Aussagen über Eigenschaften der Grundgesamtheit (alle möglichen Belegungen) treffen.
- Entwicklung von Schätz- und Testverfahren für solche Aussagen, z.B.:
  - Schätzer für Parameter von Verteilungsfunktionen.
  - Signifikanztests für Aussagen.

Numerik

- Ziel: Konstruktion und Analyse von Algorithmen für kontinuierliche mathematische Probleme, falls
  - Keine exakte Lösung für ein Problem existiert,
  - Exakte Lösung nicht effizient gefunden werden kann.
- Konstruktionsprinzipien:
  - Exakte Verfahren: exakte Lösung bei unendlicher Rechnergenauigkeit.
  - Näherungsverfahren: approximative Lösung.
- Analysen:
  - Laufzeit, Stabilität/Fehleranalyse und Robustheit.

Numerik: Fehler

- Fehlerarten:
  - Eingabefehler, Messfehler, Rundung auf Maschinengenauigkeit.
  - Systematische Fehler (z.B. Diskretisierung), Rundungsfehler.
- Beispiele:
  - Addition von x und y mit $|x| \gg |y|$: $10^{20} \neq 10^{-20} + 10^{20}$
  - Logarithmieren/Potenzrechnen: $40 \neq \ln\left(1 + e^{40}\right)$
  - Fehlerfortpflanzung: Summieren n ähnlich großer Zahlen

$$y = \sum_{i=1}^{n} x_i$$

$$y = f(1,n) \text{ mit } f(a,b) = f\left(a, \frac{a+b}{2}\right) + f\left(\frac{a+b}{2} + 1, b\right) \text{ und } f(a,a) = x_a$$

Numerik: Anwendungen

- Lösung linearer Gleichungssysteme.
- Interpolation/Approximation von reellen Funktionen.
- Finden von Extremwerten (Nullstellen, Minima, Maxima, Sattelpunkte, …) nichtlinearer Gleichungen.
- Numerische Differentiation/Integration.
- Anfangswert-/Randwertprobleme für Differential-gleichungen.
- Eigenwertprobleme und Matrix-Faktorisierung.

Numerik: Beispiel Nullstellenproblem

- Ziel: Finden von x^0 mit $g(x^0) = 0$.
- Newtonsches Näherungsverfahren:

$$x_{t+1}^0 = x_t^0 - \nabla_x g(x_t^0)^{-1} g(x_t^0)$$

  - Anwendung: Lösen von Optimierungsproblemen; für optimale Lösung x^* gilt $\nabla_x f(x^*) = 0 \Rightarrow g(x) = \nabla_x f(x)$:

$$x_{t+1}^* = x_t^* - \underbrace{\nabla_x^2 f(x_t^*)^{-1}}_{H(f)^{-1}} \underbrace{\nabla_x f(x_t^*)}_{grad(f)}$$

- Quasi-Newton-Verfahren:
  - Approximation von $\nabla_x g(x_t^0)^{-1}$ bzw. $H(f)^{-1}$.

Zusammenfassung

- Maschinelles Lernen ist zu einem großen Teil die *Anwendung von Mathematik* aus zahlreichen Gebieten, insbesondere der Statistik & Optimierung.

- Inhalt dieser Vorlesung ist
  - Das Verstehen und Implementieren von Algorithmen des Maschinellen Lernens.

- Inhalt dieser Vorlesung ist NICHT
  - Das Herleiten/Erklären der zugrunde liegenden Mathematik.