INTELLIGENTE DATENANALYSE IN MATLAB

Datenselektion und Datenaufbereitung

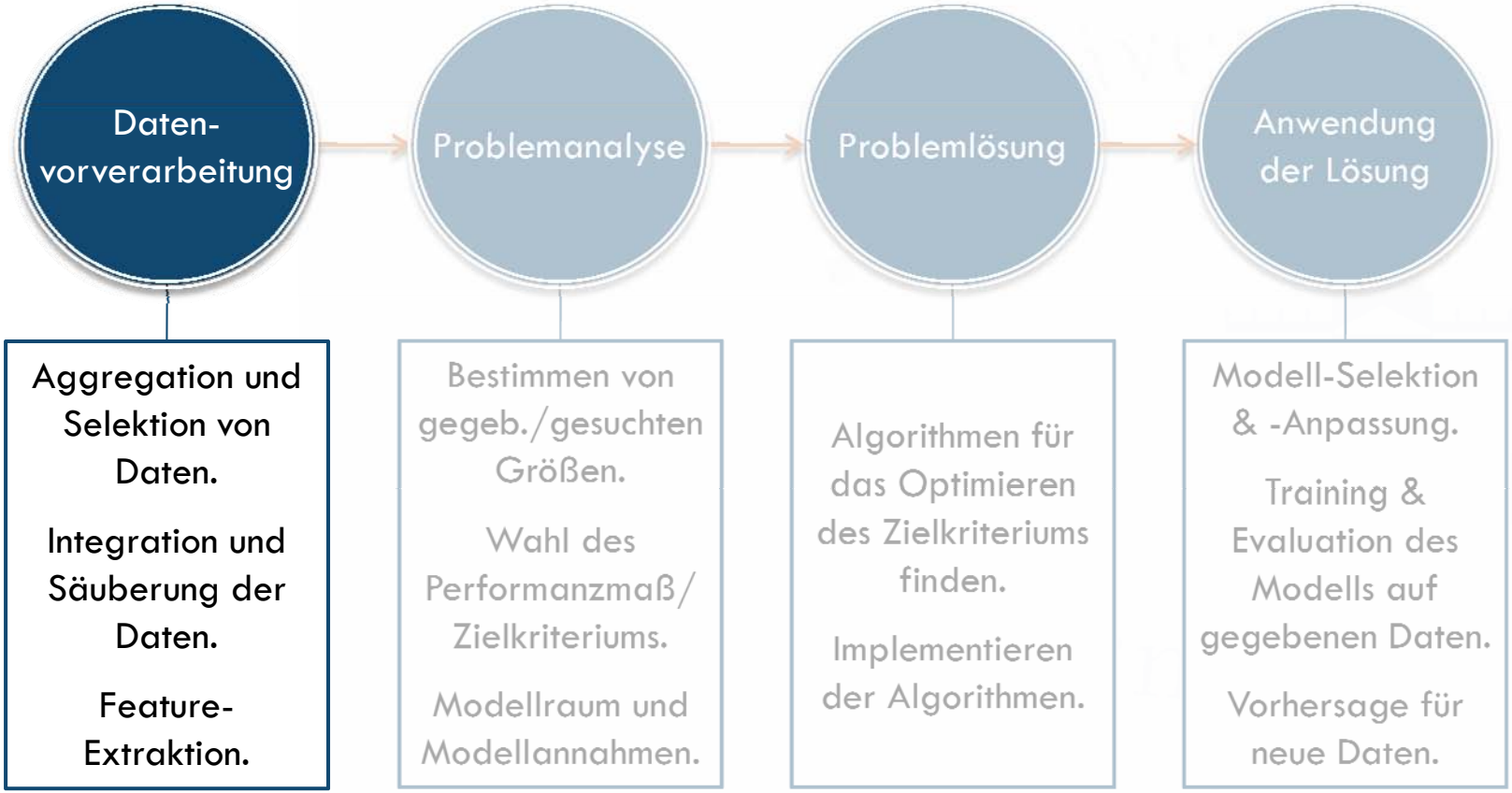
Michael Brückner/Tobias Scheffer

Literatur

- I. H. Witten, E. Frank: Data Mining – Practical Machine Learning Tools and Techniques.
- J. Han, M. Kamber: Data Mining – Concepts and Techniques.

Überblick

- Schritte der Datenanalyse:



Überblick

- Aggregation und Selektion von Daten.
  - Datenquellen identifizieren und selektieren.
  - Bestimmung von Art und Umfang der Daten.
- Integration und Säuberung der Daten.
  - Unvollständige/fehlende Daten.
  - Fehlerhafte und inkonsistente Daten.
- Feature-Extraktion.
  - Feature generieren.
  - Daten transformieren.
  - Daten- & Feature-Reduktion.

Aggregation und Selektion von Daten:

Datenquellen identifizieren und selektieren

- Genauigkeit (Anteil Rauschen in den Daten).
- Vollständigkeit (Anteil fehlender/unbrauchbarer Werte).
- Konsistenz (Widerspruch in Daten einer/mehrerer Quellen).
- Aktualität der Datenquelle.
- Glaubwürdigkeit der Datenquelle.
- Mehrwert/Redundanz.
- Verständlichkeit (Bedeutung der Attribute).
- Verfügbarkeit.

Aggregation und Selektion von Daten:

Bestimmung von Art und Umfang der Daten

- Stichprobe:
  - Auswahl repräsentativer Daten.
- Umfang der Daten:
  - Menge an unterschiedlichen Instanzen.
  - Menge an unterschiedlichen Attributen.
  - Datendichte/Sparsity.
- Art der Daten:
  - Elementare/zusammengesetzte Daten.

Aggregation und Selektion von Daten:

Bestimmung von Art und Umfang der Daten

- Elementare Datentypen:
  - Kontinuierliche Werte (z.B. Temperatur, Alter).
  - Ordinale Daten, d.h. diskrete Werte einer geordneten Menge (z.B. Schulnoten, Fahrzeugklassen, Stockwerke).
  - Nominale Daten, d.h. diskrete Werte einer ungeordneten Menge (z.B. Wörter, Familienstand, Augenfarbe, Nationalität).
- Zusammengesetzte Datentypen:
  - Tupel/Vektoren, Matrizen, Tensoren (z.B. Bilder).
  - Sequenzen (z.B. Text, EKG-Kurve, Audio-/Video-Daten).
  - Graphen, verlinkte Daten (z.B. Webseiten, Moleküle).

Integration und Säuberung der Daten

- Daten zusammentragen, konvertieren und integrieren in eine kohärente Form, z.B. in einem *Data Warehouse*.
  - Daten-Integration (z.B. unterschiedliche Speicherformate/-orte).
  - Schema-Integration (z.B. Metadaten von verschiedenen Quellen, gleiche/ähnliche Attribute unterschiedlicher Quellen)
  - Daten-Konflikte behandeln (z.B. Umrechnung von Einheiten).
  - Redundante Informationen erkennen (z.B. durch Korrelationsanalyse, Dubletten-Suche).
- Säuberung der Daten.

Integration und Säuberung der Daten:

Unvollständige/fehlende Daten

- Arten von fehlenden Daten:
  - Fehlende Werte (Attributbelegungen).
  - Fehlende Attribute welche für die Datenanalyse von Interesse wären.
  - Aggregierte Werte/Attribute.
- Ursache für fehlende Daten:
  - Zufälliges Fehlen (z.B. Speicherfehler, Fehlfunktion eines Messinstruments).
  - Systematisches Fehlen (z.B. Wert zu einem früheren Zeitpunkt unwichtig/unbekannt).
  - Daten-Integration (z.B. gelöschte Werte aufgrund von Inkonsistenzen).
  - Daten-Aggregation (z.B. aus Datenschutzgründen).
  - …

Integration und Säuberung der Daten:

Unvollständige/fehlende Daten

- Behandlung fehlender Werte:
  - Betreffende Instanzen/Attribute löschen.
  - Erweiterung des Wertebereichs (z.B. „missing“) und/oder der Attributmenge (z.B. binäres Attribut_XY_bekannt).
  - Fehlende Werte aus Daten schätzen:
    - Mean/Median Imputation (evtl. Klassen-abhängig).
    - Inferenz aus den Daten (z.B. mittels EM-Algorithmus).
  - Fehlende Werte nicht behandeln (späteres Lern-/Analyseverfahren berücksichtigt fehlende Werte).

Integration und Säuberung der Daten:

Fehlerhafte und inkonsistente Daten

- Arten von fehlerhaften Daten:
  - Rauschen in numerischen Werten. $\tilde{x} = x + \varepsilon$ $\varepsilon \sim N(0, \sigma^2)$
  - Ausreißer/falsch gelabelte Daten.
  - Inkonsistenz.
- Ursache für fehlerhafte Daten:
  - Zufällige Störung (z.B. durch Messungenauigkeit).
  - Systematische Störung (z.B. technische Grenzen, Formatierungsfehler).
  - Unabsichtliche Störung (z.B. falsches Labeling durch menschliche Fehler).
  - Absichtliche Störung (z.B. durch Spammer bei Emails).
  - Daten-Integration (z.B. fehlerhafte Konvertierung/Datenübertragung, Inkonsistenzen durch sich wiedersprechende Datenquellen).
  - …

Integration und Säuberung der Daten:

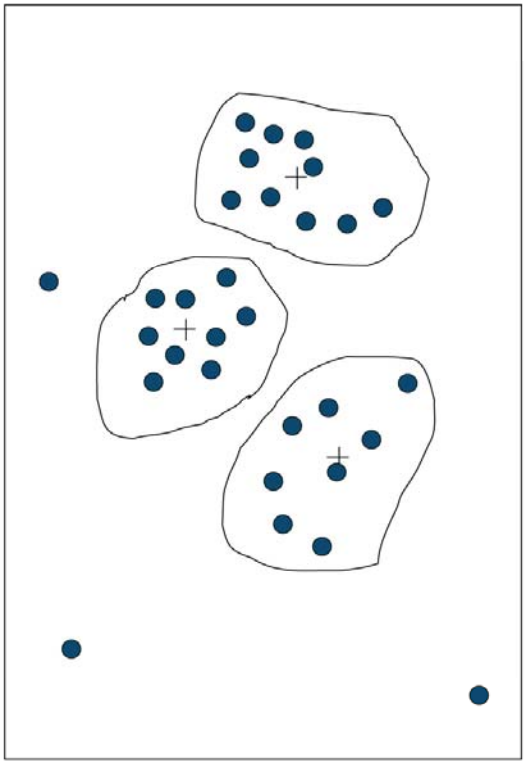
Fehlerhafte und inkonsistente Daten

- Identifizierung fehlerhafter Werte:
  - Binning: äquidistante Diskretisierung in *Bins*
    $\Rightarrow$ Bins mit einem/wenigen Instanzen enthalten evtl. Ausreißer.
  - Clustering: Suche nach Regionen mit hoher Datendichte (*Cluster*)
    $\Rightarrow$ Cluster mit einem/wenigen Instanzen enthalten evtl. Ausreißer.
  - Active Learning: Wiederspruch zwischen Daten und Modell
    $\Rightarrow$ auffällige Instanzen werden Menschen zur Kontrolle vorgeschlagen.
- Behandlung fehlerhafter Werte:
  - Glättung numerischer Werte (z.B. Regression, Moving Average).
  - Diskretisierung (z.B. Alter $\Rightarrow$ {kücken, bivi, uhu}).
  - Als fehlend behandeln.

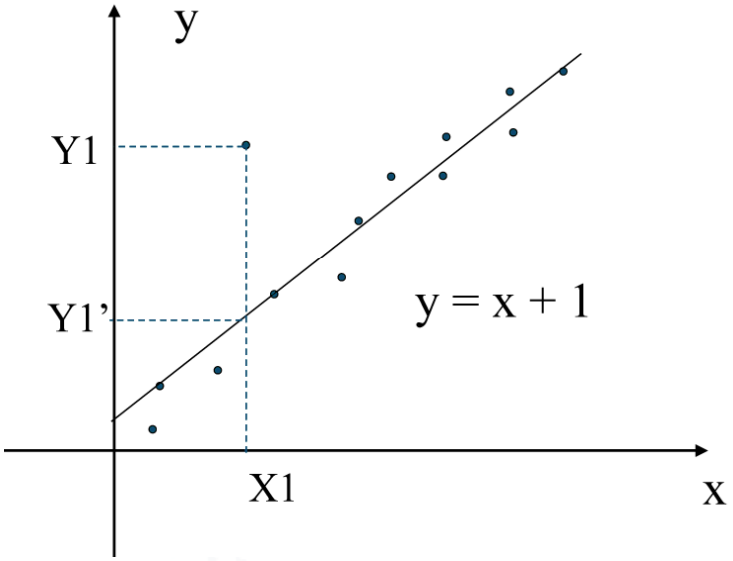
Integration und Säuberung der Daten:

Fehlerhafte und inkonsistente Daten

Clustering



Lineare Regression



Feature-Extraktion:

Feature generieren

- Text-Vorverarbeitung:
  - Konvertierung (z.B. Zeichensatz, Klein-/Großschreibung, HTML-Decoding).
  - Tokenisierung.
  - Stopwords (z.B. Artikel, Konjunktionen) entfernen.
  - Stemming (z.B. Verben in Infinitivform, Substantive in Singularform).
- Feature aus Texten:
  - TF-IDF (Term Frequency – Inverse Document Frequency).
  - Bag-of-Words (Wörter sind nominale bzw. binäre Attribute).
  - *N*-Gram (Wort-Tupel der Länge *N* sind nominale bzw. binäre Attribute).
  - OSB (Orthogonal Sparse Bigrams).
  - Darstellung als Baum/Graph (z.B. bei XML).

Feature-Extraktion:

Feature generieren

- Bag-of-Words:

„Hello World“ $\Rightarrow [0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0]^T$

Index von „World“

Index von „Hello“

- TF-IDF:

$$tf_{i,j} = \frac{freq_{i,j}}{\max_k \left(freq_{k,j} \right)}$$

$$idf_i = \log \frac{N}{n_i}$$

$$w_{i,j} = tf_{i,j} \cdot idf_i = \frac{freq_{i,j}}{\max_k \left(freq_{k,j} \right)} \log \frac{N}{n_i}$$

Feature-Extraktion:

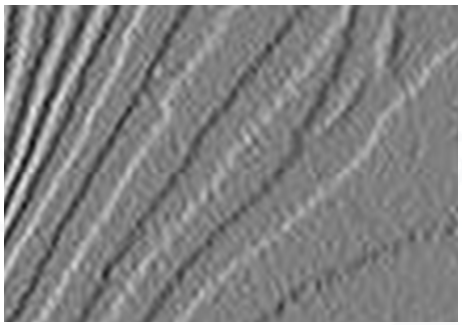
Feature generieren

- Bild-Vorverarbeitung:
  - Normierung (z.B. Farbwerte, Helligkeit, Farbintensität, Größe, Lage).
  - Transformation (z.B. Skalierung, Verschiebung, Drehung).
  - Filtern (z.B. Glättung, Kantenhervorhebung).
- Feature aus Bildern:
  - Metainformationen (z.B. Tags, Format, Größe, Farbverteilung).
  - Pixelvektor (Umwandlung der Bildmatrix in Vektor).
  - FFT (Fast Fourier Transform).
  - SIFT (Scale-Invariant Feature Transform).
  - HT (Hough Transform).

Feature-Extraktion:

Feature generieren

Convolution Kernel



Feature-Extraktion:

Feature generieren

- Vorverarbeitung verlinkter Daten:
  - Umwandlung verlinkter Daten in (un-)gewicheten, (un-)gerichten Graph.
  - Graph-Konvertierung (z.B. Minimal Spanning Tree/Forest, zusammenhängender Graph, Zyklen-freier Graph).
- Feature aus verlinkten Daten:
  - Knoten-Attribute (z.B. Anzahl ein-/ausgehender Kanten).
  - Kanten-Attribute (z.B. verbindende Knoten, Gewicht, Richtung).
  - Häufige Subgraphen, Subtrees, Cliquen.

Feature-Extraktion:

Daten transformieren

- Feature-Normalisierung:
  - Min/Max-Normalisierung: $$x^{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \left(x^{new}_{\max} - x^{new}_{\min} \right) + x^{new}_{\min}$$
  - Z-Score-Normalisierung: $$x^{new} = \frac{x - \mu_x}{\sigma_x}$$
  - Dezimal-Skalierung: $$x^{new} = |x| \cdot 10^a \qquad a = \max \left\{ i \in \mathbb{Z} \,|\, |x| \cdot 10^i < 1 \right\}$$
  - Logarithmische Skalierung: $$x^{new} = \log_a x$$

Feature-Extraktion:

Daten transformieren

- Instanzen normalisieren:
  - Unabhängige l_a-Normierung aller Datenpunkte: $\mathbf{x}^{new} = \|\mathbf{x}\|_a^{-1} \mathbf{x}$
  - Korrelation zw. Attributen entfernen: $\mathbf{x}^{new} \sim N(\mathbf{0}, \mathbf{I})$
    - LU-Zerlegung der Kovarianz-Matrix $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}$.
    - Transformation der Eingaben: $\mathbf{x}^{new} = \mathbf{L}^{-1/2}\mathbf{U}^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}})$
- Feature-Konstruktion:
  - Kombination elementarer Feature (z.B. $(x_i, x_j) \rightarrow \left(x_j, \sqrt{x_i x_j}, x_i + x_j\right)$).
  - Mapping elementarer Feature (z.B. $x_i \rightarrow \left(x_i, \log x_i, x_i^2\right)$).

Feature-Extraktion:

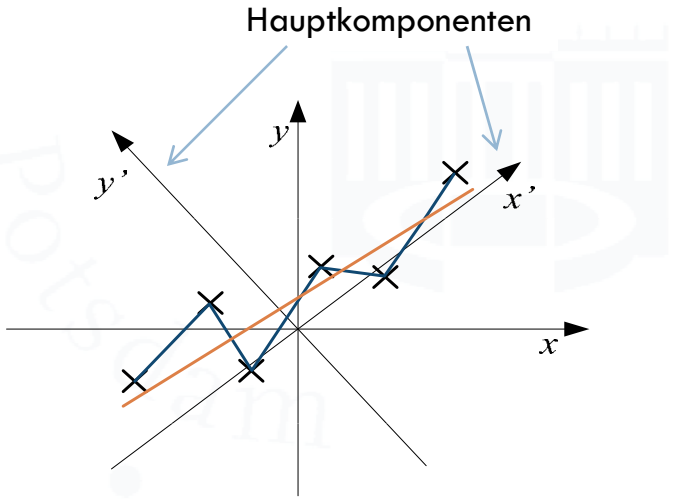
Daten transformieren

- Diskretisierung:
  - Binning/Clustering: Wertebereich auf Mittelpunkte/Mediane der Bins bzw. Cluster einschränken.
  - Entropie-basierte Diskretisierung:
    Partitionierung der Menge S in S_1 und S_2 durch $\arg\min_{S=S_1 \cup S_2} \frac{|S_1|}{|S|} H_{S_1} + \frac{|S_2|}{|S|} H_{S_2}$.
  - Binärisierung:
    - Numerische Werte (durch Schwellwert).
    - Nominale Werte (jeder Wert des Wertebereichs wird ein eigenes, binäres Attribut).
  - Natürliche Partitionierung (z.B. durch Hierarchien/Taxonomien).

Feature-Extraktion:

Daten- & Feature-Reduktion

- Feature-Selektion (Dimensionsreduktion):
  - Vorwärts-Selektion bzw. Rückwärts-Elimination (Kombination aus beiden).
  - Elimination nach dem Lernen.
  - Automatische Selektion beim Lernen (z.B. Entscheidungsbaumlernen).
- Datenkompression:
  - Wavelet-/Fourier-Transformation.
  - Hauptkomponenten-/Faktoranalyse.
  - Interpolation (z.B. lineare Interpolation, Spline-Interpolation).
  - Regression (z.B. lineare Regression).
  - Sampling.



Zusammenfassung

- Datenvorverarbeitung entscheidend für Qualität der Datenanalyse.
- Dient der Bereitstellung aller *relevanten* Daten in der *nützlichsten* Form.
- Sie umfasst
  - Aggregation und Selektion von Daten,
  - Integration und Säuberung der Daten und
  - Feature-Extraktion.
- Vorverarbeitung oft zeitaufwendiger als Datenanalyse!