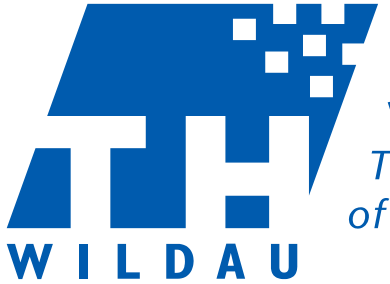

Technische
Hochschule
Wildau [FH]
Technical University
of Applied Sciences

Maschinelles Lernen

Datenvorverarbeitung

Prof. Dr. Rainer Stollhoff

Univariate Vorverarbeitung - Datentypkonvertierung

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

- Nominelle Variablen
  - Character <-> Factor
    - Für viele Verfahren nicht notwendig, da automatisch angewandt
  - Dummy-Kodierung (One-Off-Kodierung)
    - Übersetzt nominalen Faktor mit n Werten in n separate 0/1-Variablen (Variante n-1 Variablen)
    - Bei manchen Verfahren notwendige Vorverarbeitung, bei anderen Verfahren integriert
- Numerische Variablen
  - Zahlenwert als Faktor
    - Notwendig: falls Klassifikationsaufgabe und Klassenkodierung als Zahlenwert – sonst automatisch Regressionsverfahren
    - Ermöglicht Einsatz von Klassifikationsverfahren in Regressionsproblemen
    - Diskretisierung z.B. zur Visualisierung als Histogramm
  - Faktor als Zahlenwert
    - Ermöglicht Einsatz von Regressionsverfahren für Klassifikationsaufgaben – aber Vorsicht: nur sinnvoll für ordinale Merkmale!

Univariate Vorverarbeitung – Imputation fehlender Werte

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

Mittelwert / Median

- Ersetzt für eine Beobachtung fehlende Werte in einer Variable durch den Mittelwert bzw. Median dieser Variable in anderen Beobachtungen

- Vorteile
  - Einfach und Robust
  - Alle Beobachtungen können verwendet werden
- Nachteile
  - Ignoriert Zusammenhänge zwischen Variablen

(lokal)-lineare Modell

- Schätzt fehlende Werte in einer Variablen anhand eines Regressionsmodells, das auf allen anderen Variablen geschätzt wird

- Vorteile
  - Berücksichtigt Zusammenhänge zwischen Variablen
- Nachteile
  - Zusätzlicher Rechenaufwand

Univariate Vorverarbeitung - Transformation

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

Standardisierung der Variablen

- Abziehen des Mittelwerts und teilen durch die Standardabweichung

$$z = \frac{x - \bar{x}}{\sigma(x)}$$

- Normalisierung des Wertebereichs auf [0,1]

$$\check{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Vorteile
  - Eingangsvariablen in Modelle haben vergleichbare Skala / Auflösung
  - Parameter in einem Regressionsmodell können direkt verglichen werden
- Nachteile
  - Einheiten gehen verloren
  - Zusätzlicher Rechenaufwand

Nicht-lineare Zusammenhänge aufnehmen

- Beispiele:
  - Quadratischer Zusammenhang x^2
  - Polynomieller Zusammenhang x^n
  - Betragsmäßiger Zusammenhang $|x|$
  - Logarithmische Skalierung $\ln(x)$
- Vorteile
  - Abbildung von Vorkenntnissen z.B. physikalischer Gesetzmäßigkeiten $E = \frac{1}{2}mv^2$
  - Erweiterung linearer Verfahren z.B. lineare Regression
- Nachteile
  - Zusätzlicher Rechenaufwand
  - Kein Vorteil bei Verfahren mit eingebauter Transformation bzw. Unabhängigkeit ggb. Transformationen z.B. Bäume/rekursive Partitionierung

Multivariate Vorverarbeitung

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

Dimensionsreduktion

(vgl. Unsupervised Learning)

- **Korrelationen herausnehmen**
  - Hauptkomponentenanalyse durchführen
  - Statt Variablen Hauptkomponenten verwenden
- **Embeddings**
  - Vielzahl von Dummy-Variablen in metrischen Raum einbetten (z.B. Worträume)

- Vorteile
  - Geringere Speicherbelegung
  - Notwendig in linearer Regression, falls Ausgangsvariablen linear abhängig
- Nachteile
  - U.U. Verlust von Informationen
  - Verfahren sind datengetrieben

Interaktionen berechnen

- Beispiele:
  - Zweifaches Produkt $x_1 \cdot x_2$
  - mehrfaches Produkt $x_1 \cdot x_2 \cdot x_3 \cdots$
  - Exponent $x_1^{x_2}$
  - Quadratischer Abstand $(x_1 - x_2)^2$
- Vorteile
  - Abbildung von Vorkenntnissen z.B. physikalischer Gesetzmäßigkeiten $E = \frac{1}{2}mv^2$
  - Erweiterung linearer Verfahren z.B. lineare Regression
- Nachteile
  - Zusätzlicher Rechenaufwand
  - Kein Vorteil bei Verfahren mit eingebauter Interaktionsmöglichkeit z.B. Bäume/rekursive Partitionierung