Kurs: KI T-InfT-008 und 010 Datenmengen und Embedded Systems

Cândido Vieira

26.09.2024

Balthasar-Neumann-Technikum (BNT)

Inhaltsverzeichnis

1. Einführung Datenbereinigung

- **Datenqualität beeinflusst Modellqualität**: Fehlerhafte oder unvollständige Daten führen zu falschen Vorhersagen.

- **80% der Zeit in Datenaufbereitung**: Laut Studien verbringen Datenwissenschaftler den Großteil der Zeit mit der Bereinigung und Vorbereitung von Daten.

- **Schlechte Daten = Schlechte Modelle**: Auch der beste Algorithmus kann keine fehlerhaften oder unvollständigen Daten korrigieren.

1. Datenbereinigung - Häufige Datenprobleme

- **Fehlende Werte**: Zellen ohne Datenpunkte.
- **Fehlerhafte Daten**: Daten, die unrealistische oder falsche Werte enthalten.
- **Duplikate**: Doppelte Einträge, die die Analyse verzerren können.
- **Ausreißer**: Extremwerte, die nicht zur Datenverteilung passen.

2. Methoden - Umgang mit fehlenden Daten (1/2)

- **Entfernen von Datenpunkten**:
  - Entfernen der gesamten Zeile oder Spalte, in der Daten fehlen.
  - **Vorteil**: Einfach anzuwenden.
  - **Nachteil**: Verlust wertvoller Daten.

- **Auffüllen mit Standardwerten**:
  - Auffüllen mit Werten wie **0, Mittelwert, Median** oder **Modus**.
  - **Vorteil**: Einfach und schnell.
  - **Nachteil**: Kann Verzerrungen einführen.

2. Methoden - Umgang mit fehlenden Daten (2/2)

- **Interpolation**:
  - Nutzen von benachbarten Datenpunkten, um den fehlenden Wert zu schätzen.
  - **Vorteil**: Nützlich bei Zeitreihendaten.

- **Vorhersagemodelle**:
  - Modelle wie **KNN (K-Nearest Neighbors)** zur Schätzung fehlender Werte.
  - **Vorteil**: Präzisere Schätzungen.
  - **Nachteil**: Rechenaufwendig.

2. Methoden - Umgang mit fehlerhaften Daten

- **Fehlerhafte Daten entfernen oder korrigieren**:
  - **Beispiel**: Unrealistische Werte wie negative Körpergröße.

- **Duplikate entfernen**:
  - Doppelte Zeilen oder Einträge, die durch Mehrfacheingaben entstehen.
  - pandas: .drop_duplicates()

- **Ausreißer erkennen und behandeln**:
  - Identifikation durch statistische Methoden (z.B. Z-Score, IQR).

3. Einführung - pandas für die Datenbereinigung (1/2)

- **pandas**: ist eine Python-Bibliothek, die für Datenaufbereitung und -analyse verwendet wird.

- Die **zentrale Datenstruktur** ist der **DataFrame**, eine tabellenähnliche Struktur, die Spalten und Zeilen enthält.

3. Einführung - pandas für die Datenbereinigung (2/2)

- **Wichtige Funktionen in pandas**:
  - 1. **Laden von Daten**: pd.read_csv("datei.csv")
  - 2. **Daten anzeigen**: .head(), .tail()
  - 3. **Fehlende Werte erkennen**: .isna(), .isnull()
  - 4. **Daten bereinigen**: - **Fehlende Werte entfernen**: .dropna() - **Auffüllen von Werten**: .fillna(value)
  - 5. **Duplikate entfernen**: .drop_duplicates()

4. **Implementierung**

Beispiel für Datenbereinigung mit pandas:

```
import pandas as pd

# CSV-Datei laden
df = pd.read_csv("daten.csv")

# Fehlende Werte anzeigen
print(df.isna().sum())

# Fehlende Werte mit Median auffüllen
df['Spalte'] = df['Spalte'].fillna(df['Spalte'].median())

# Duplikate entfernen
df = df.drop_duplicates()

# Ergebnis speichern
df.to_csv("bereinigte_daten.csv")
```

4. **Implementierung**

- **Schritt 1**:
  - Daten laden und erkunden: .head(), .describe()
- **Schritt 2**:
  - Fehlende Werte identifizieren und behandeln: .isna(), .fillna()
- **Schritt 3**:
  - Duplikate und Ausreißer finden und beseitigen: .drop_duplicates()
- **Schritt 4**:
  - Bereinigte Daten speichern :.to_csv()

5. Zusammenfassung der pandas-Funktionen

Diese Übersicht fasst die wesentlichen **pandas**-Funktionen zusammen, die für die Datenbereinigung notwendig sind.

df = pd.read_csv(“datei.csv”)	Daten laden	Lädt Daten aus einer CSV-Datei in einen DataFrame
df.head()	Daten anzeigen	Zeigt die ersten 5 Zeilen des DataFrames an
df.describe()	Zusammenfassung anzeigen	Gibt eine statistische Übersicht der numerischen Spalten
df.isna().sum()	Fehlende Werte anzeigen	Zeigt die Anzahl der fehlenden Werte in jeder Spalte
df.dropna()	Zeilen mit fehlenden Werten entfernen	Entfernt alle Zeilen, die fehlende Werte enthalten
df[‘Spalte’].fillna(df[‘Spalte’].mean())	Fehlende Werte auffüllen	Füllt fehlende Werte in einer Spalte mit dem Mittelwert auf
df.drop_duplicates()	Duplikate entfernen	Entfernt doppelte Zeilen aus dem DataFrame
df.to_csv(“bereinigte_daten.csv”, index=False)	DataFrame speichern	Speichert den bereinigten DataFrame in einer CSV-Datei

6. Zusammenfassung

- **Datenbereinigung ist entscheidend**: Schlechte Daten führen zu schlechten Modellen.

- **Fehlende Daten** können entweder entfernt, interpoliert oder durch Vorhersagen gefüllt werden.

- **pandas** bietet leistungsstarke Funktionen für die Datenbereinigung und analyse.

7. Fragen & Diskussion

- Welche Herausforderungen haben Sie bei der Datenbereinigung gesehen?
- Fragen zur praktischen Anwendung von pandas?