Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Unüberwachtes Lernen: Feature Engineering Praxis

Violeta Vogel

Technik und Informatik, BFH

Aufbau eines Data Frames

- Objekte / Beobachtungen sind in Zeilen (rows) angeordnet
- Merkmale / Attribute (Variablen, Features) sind in Spalten (columns) angeordnet
  - Spalten enthalten (idealerweise) sprechende Namen, über welche sie angesprochen werden können
  - pro Spalte ist ein Datentyp festgelegt, unterschiedliche Spalten können aber unterschiedliche Typen aufweisen
- analog einer Tabelle in einer Relationalen Datenbank
- pandas.DataFrame: der Objekttyp, mit welchem in Folgenden hauptsächlich gearbeitet wird: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html

Aufbau eines Data Frames

- die für diesen Kurs vorbereiteten und abgegebenen Beispieldaten werden ausschliesslich im CSV-Format zur Verfügung gestellt
- die ersten Schritte sind daher
  - die Daten lokal hinterlegen (Filesystem)
  - den Datenpfad festlegen
  - Das Dataset (CSV) in ein (pandas.) DataFrame Objekt einlesen
- die hier vorgestellten Methoden beziehen die Daten danach ausschliesslich von geladenen DataFrame Objekten

Beispieldaten: Lehrbuchbeispiele

- viele Methoden werden mit idealisierten Datenbeispielen eingeführt
- man kann davon ausgehen, dass dort das Preprocessing (Feature Engineering) bereits erfolgt ist
- ausserdem sind sie derart optimiert, dass die entsprechenden Methoden verständlich visualisiert werden können
  - oft nur zwei Variablen
  - Target bei Klassifikationsbeispielen nur zwei Klassen
  - eine Modellvalidierung wird bei diesen Beispielen nicht vorgenommen, sie dienen primär dem Aufzeigen der jeweils zugrunde liegenden Lernalgorithmen

Beispieldaten: Fallstudie 1 Klassifikation

- Fallstudie 1: Bank Marketing Data Set (https://archive.ics.uci.edu/dataset/222/bank+marketing)
- stammt aus einer Marketing Campagne eines portugiesischen Bankinstituts
- das Ziel der Klassifikation ist es, vorherzusagen, ob sich ein Kunde für ein bestimmtes Anlageprodukt überzeugen lässt oder nicht
- im Original (im hinterlegten Link unter bank-additional-full.csv) besteht das Dataset aus 45211 rows und 21 columns
- aus praktischen Gründen wurde das Dataset für diesen Kurs etwas modifiziert
- dieses soll im Rahmen von Feature Engineering für das anschliessende Modellieren in
- Supervised Learning (Klassifikation) gemäss dessen Anforderungen aufbereitet werden
- Dateiname: bank_data.csv (auf Moodle)
- Target: "y"
- Delimiter: ","

Beispieldaten: Fallstudie 2 Regression

- für Überwaches Lernen - Regression wird eine andere Fallstudie eingesetzt: Immobiliendaten von Melbourne, https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market
- diese Daten wurden aus öffentlich zugänglichen Ergebnissen zusammengestellt, die jede Woche von Domain.com.au veröffentlicht werden
- der Datensatz enthält Adresse, Immobilientyp, Quartier, Verkaufsmethode, Anzahl Räume, Preis, Immobilienmakler, Verkaufsdatum und Entfernung von C.B.D. (Central Business District, Geschäftsviertel von Melbourne)
- Dateiname: melb_data.csv (auf Moodle)
- Target: "Price"
- Delimiter: ";"

Beispieldaten

- Daten werden entweder als CSV-Dateien abgegeben

```
import pandas as pd
demo_data_class = pd.read_csv('demo_data_class.csv')
demo_data_regr = pd.read_csv('demo_data_regr.csv')
```

- Oder können direkt aus Packages geladen werden. Z.B.: Aus Seaborn

```
import seaborn as sns
iris_data = sns.load_dataset('iris')
```

Python Libraries

- alle abgegebenen Code Beispiele basieren auf
- Anaconda3-2024.02-1-Windows-x86_64 Distribution
- Python 3.11.7
- und den folgenden Libraries:

Library	Alias	Beschreibung	Vers.
numpy	np	Programmbibliothek für die Programmiersprache Python, die eine einfache Handhabung von Vektoren, Matrizen oder generell grossen mehrdimensionalen Arrays ermöglicht	1.26.4
pandas	pd	eine Programmbibliothek für die Programmiersprache Python, die Hilfsmittel für die Verwaltung von Daten und deren Analyse anbietet	2.1.4
matplotlib.pyplot	plt	mathematische Visualisierungen aller Art, die Synthax basiert auf Mathlab	3.7.5

Python Libraries

Library	Alias	Beschreibung	Vers.
seaborn	sns	basierend auf matplotlib zum Erstellen von attraktiven und informativen statistischen Visualisierungen	0.12.2
pandas-profiling (ydata_profiling)	---	erstellt Profiling Reports, ausgehend von pandas DataFrames; erweitert pandas DataFrame mit der Methode ProfileReport () für schnelle Datenanalysen	4.8.3
scikit-learn (sklearn)	---	die wichtigste Programmbibliothek für den Kursteil Supervised Learning	1.4.2
imbalanced-learn (imblearn)	---	eine Library für den Umgang mit unbalancierten Daten	0.12.2
statsmodels	stat	ein Modul mit Funktionen und Klassen für klassisch-statistische Aufgabenstellungen (optional)	0.14.0

Workshop: EDA Skalierung

- Gruppen zu 3 bis 4, Zeit: 30'
- untersuchen Sie die Variablen des Melbourne Housing Dataset (melb_data.csv) auf vorliegende Skalenniveaus:
  - kategorial (auch Unterscheidung nominal / ordinal)
  - metrisch
- erstellen Sie eine tabellarische Zusammenstellung mit einem Tool Ihrer Wahl derart, welche sie in folgenden Workshops mit weiteren Informationen ergänzen können (in der vorbereiteten Tabelle auf MS Teams können die Ergebnisse für die ganze Klasse konsolidiert hinterlegt werden)
- konsultieren Sie auch die Online-Dokumentation (ist sie zu den Daten konsistent?)
- falls Sie die Daten gleich mit Python sichten möchten, können diese mit untenstehendem Code geladen werden (*der Pfad muss natürlich individuell angepasst werden*)

```
import pandas as pd
pd.read_csv('../3_data/melb_data.csv')
```