Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

2024 FS CAS PML - Supervised Learning
3 Regression
3.1 Einführung

Werner Dähler 2024

3 Regression - AGENDA

31. Einleitung
    - 311. Abgrenzung
    - 312. Demo Dataset
    - 313. Fallstudien Dataset
    - 314. Vorbereiten der Umgebung
32. Regression klassisch (OLS)
33. Regression mit ML
34. Vergleiche über alle Modelle

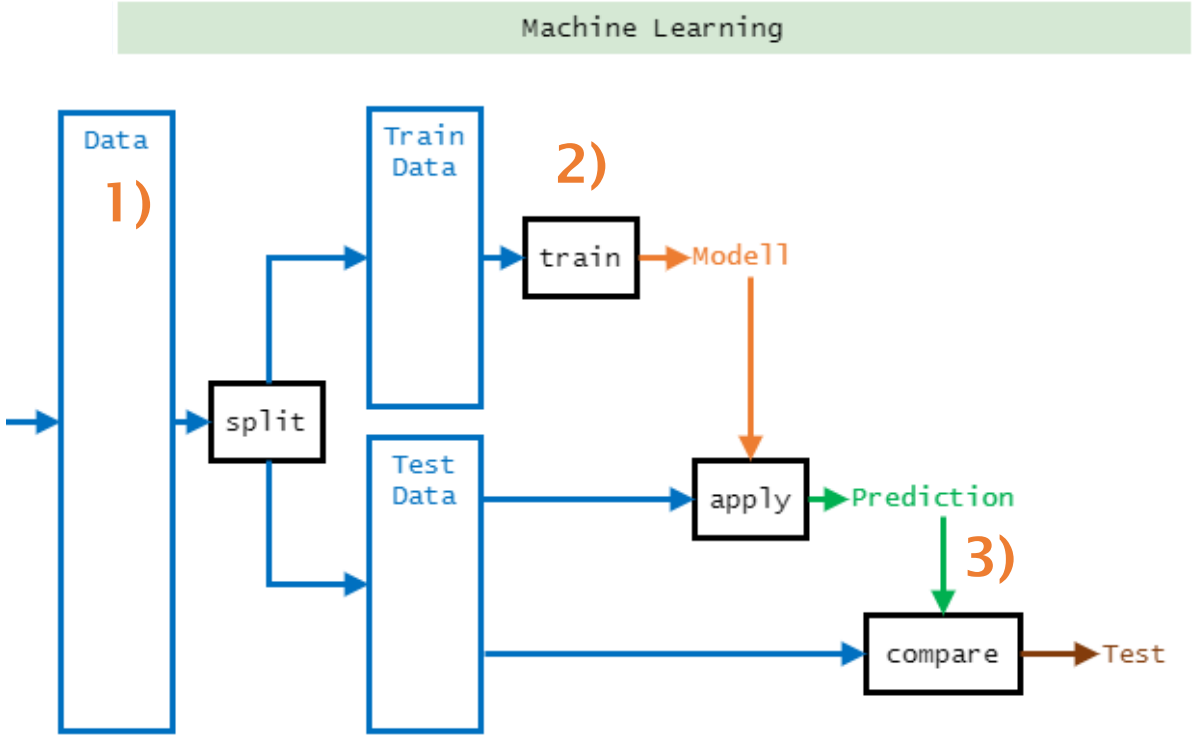
die hinterlegten Links wurden am 27.02.2024 abgegriffen

3.1 Regression - Einleitung

3.1.1 Abgrenzungen gegenüber Klassifikation

1) in den Daten: das Target ist **metrisch** skaliert und weist (meist) stetige numerische Werte auf
2) andere Vorhersageklassen: zumeist `xxxRegressor` anstelle von `xxxClassifier`
3) andere Performance-Metriken (z.B. r2_score anstelle von accuracy, Details dazu in Kap. 4.4.2)

- im Übrigen ist das Vorgehen dasselbe wie bei Klassifikation

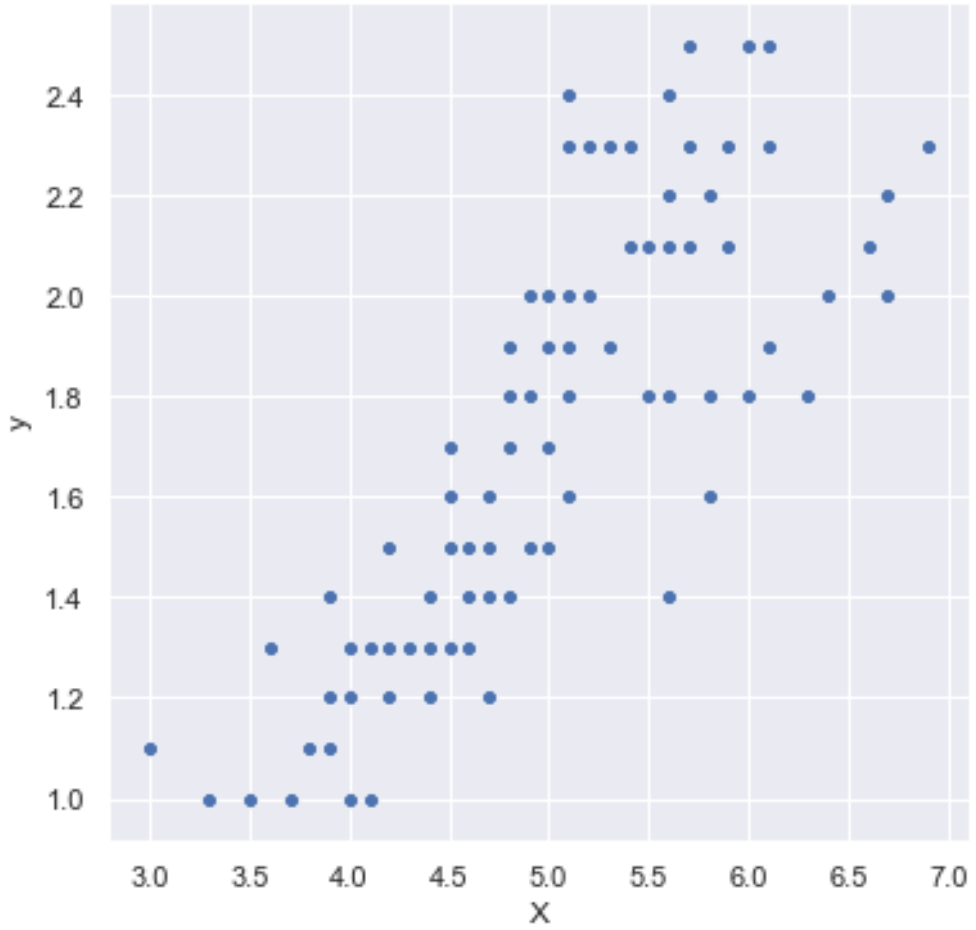


3.1 Regression - Einleitung

3.1.2 Das Demo Dataset

- zur Visualisierung von Ideen und Verfahren wird auch hier ein Demo Dataset verwendet:
  - demo_data_regr.csv
- zwei Spalten (columns)
  - X: Feature (unabhängige Variable)
  - y:Target (abhängige Variable)
- 81 Beobachtungen

- (es sind tatsächlich dieselben Daten wie bei demo_data_class.csv, ausser
  - X1 $\rightarrow$ X
  - X2 $\rightarrow$ y)



3.1 Regression - Einleitung

3.1.3 Das Fallstudien Dataset

- wie bei der Klassifikation wird auch bei Regression mit einem Dataset aus einer konkreten Fallstudie gearbeitet
- das für die Praxisteile im Rahmen von Regression verwendete Dataset wurde im Rahmen des Workshops 03 unter Feature Engineering bereits aufbereitet
- einige Kennwerte
  - Anzahl rows: 18'393
  - Anzahl columns: 24,
    davon
    - float64: 10
    - int64:14
  - Target: "Price" (float64)
- Ziel der Arbeiten mit diesem Dataset: trainieren eines Vorhersagemodells für den Verkaufspreis von Immobilien

3.1 Regression - Einleitung

3.1.4 Vorbereiten der Umgebung

- wie bei den Methoden zur Klassifikation hat es auch bei der Regression im begleitenden Jupyter Notebook gleich am Anfang einen Codeblock, in welchem die Umgebung und die Daten vorbereitet werden:
  - importieren der notwendigen Libraries
  - setzen des Datenpfades
  - Laden und vorbereiten der Datasets
    - Demo Dataset
    - Melbourne Housing Dataset

- das Demo Dataset wird auch hier **nicht** in Train - Test gesplittet, es wird ausschliesslich dazu verwendet, die Regressionsmethoden darzustellen (X_demo, y_demo)
- die Performance Vergleiche erfolgen dann aber auf dem Melbourne Housing Dataset, welches aus diesem Grund gesplittet wird (X_train, y_train, X_test, y_test)

3.1 Regression - Einleitung

3.1.4 Vorbereiten der Umgebung

- die Standard Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

- analog dem Vorgehen bei Klassifikation können auch hier die im Modul bfh_cas_pml implementierten Funktionen für die Bereitstellung der Daten verwendet werden

```
from bfh_cas_pml import prep_data, prep_demo_data
X_train, X_test, y_train, y_test = prep_data(
    'melb_data_prep.csv', 'Price', seed = 1234)
X_demo, y_demo = prep_demo_data('demo_data_regr.csv', 'y')
```