

Data Processing and Machine Learning WS

2024/25

Exercise 6

Maschinelles Lernen

Maschinelles Lernen ist eine Methode, bei der Systeme aus Daten lernen, um Entscheidungen zu treffen oder Vorhersagen zu machen.

Aufgabe 1: Spam-Filter mit Wortbox

- i. Entwickeln Sie ein Modell, das Nachrichten in "Spam" und "Nicht-Spam" klassifiziert.
- ii. Verwenden Sie dazu eine Wortbox, in die Wörter wie "lottery" oder "today" eingegeben werden können, um herauszufinden, ob eine Nachricht als Spam klassifiziert wird.

Aufgabe 2: Korrelationen und PCA im Boston-Datensatz

- i. Analysieren Sie die Korrelationen zwischen den Variablen im Boston-Datensatz und plotten Sie eine Heatmap, um die Korrelationswerte zwischen den Variablen zu veranschaulichen.
- ii. Identifizieren Sie die Variablen mit der höchsten Korrelation zu "MEDV".
- iii. Standardisieren Sie die Daten mit StandardScaler.
- iv. Wenden Sie PCA (Hauptkomponentenanalyse) auf den Datensatz an und plotten Sie die Korrelationsmatrix sowohl für den ursprünglichen Datensatz als auch für den PCA-transformierten Datensatz.
- v. Interpretieren Sie die Ergebnisse der PCA.

Aufgabe3: Vorhersage des MEDV mit Polynomialer Regression

- i. Standardisieren Sie die Daten mit StandardScaler.
- ii. Wenden Sie eine polynomiale Regression auf den Boston-Datensatz an, um "MEDV" vorherzusagen.
- iii. Verwenden Sie verschiedene Grade des Polynoms und vergleichen Sie die Ergebnisse.
- iv. Plotten Sie die Modelle und berechnen Sie die R^2 - und MAE-Werte.

Aufgabe 4: Vorhersage des MEDV mit Random Forest

- i. Standardisieren Sie die Daten mit StandardScaler.
- ii. Erstellen Sie ein Random-Forest-Modell, um "MEDV" im Boston-Datensatz vorherzusagen.
- iii. Führen Sie dabei eine Kreuzvalidierung mit 5 Folds durch, um die Leistung des Modells zu bewerten.
- iv. Plotten Sie die Ergebnisse und berechnen Sie die R^2 - und MAE-Werte.