Data Processing and Machine Learning WS

2024/25

Exercise 5

Maschinelles Lernen

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

Aufgabe 1: Kartoffelproduktion in Deutschland

i. Laden Sie den Datensatz und filtern Sie ihn nach „Germany“ und „Potatoes Production“ ab dem Jahr 2003.
ii. Verarbeiten Sie die Daten vor, indem Sie StandardScaler verwenden, und trainieren Sie ein Linear Regression model.
iii. Bewerten Sie das Modell mithilfe von MAE und R^2.
iv. Prognostizieren Sie die Potatoes Production für das Jahr 2030.
v. Visualisieren Sie die Originaldaten, die Regressionslinie und die Prognose für 2030

Aufgabe 2: Hühnerfleischproduktion in den Vereinigten Staaten

i. Laden Sie den Datensatz und filtern Sie ihn nach „United States“ und „Meat, chicken Production“ ab dem Jahr 2000.
ii. Verarbeiten Sie die Daten vor und trainieren Sie ein Linear Regression model.
iii. Bewerten Sie das Modell mithilfe von MAE und R^2.
iv. Prognostizieren Sie die Produktion für die Jahre 2025–2030.
v. Visualisieren Sie die Originaldaten, die Regressionslinie und die zukünftigen Prognosen.

Aufgabe3: Maisproduktion in Mexiko (SVR-Analyse)

i. Laden Sie den Datensatz und filtern Sie ihn nach „Mexico“ und „Maize Production“ (ohne Einschränkung auf ein bestimmtes Jahr).
ii. Verarbeiten Sie die Daten vor, indem Sie StandardScaler verwenden.
iii. Trainieren Sie zwei SVR-Modelle:
iv. Linear Kernel: Für eine lineare Beziehung.
v. RBF-Kernel: Um potenzielle nichtlineare Muster zu erfassen.
vi. Bewerten Sie beide Modelle mithilfe von MAE und R^2.
vii. Prognostizieren Sie die Maisproduktion für die Jahre 2025–2030 mit beiden Modellen.
viii. Visualisieren Sie:

- Originaldatenpunkte.
- Regressionslinien für beide Modelle.
- Prognosen für 2025–2030 aus beiden Modellen.

Aufgabe 4: Optimierung im überwachten Lernen

Im überwachten Lernen beinhaltet das Optimierungsproblem häufig die Minimierung einer Kostenfunktion. Erklären Sie, wie der Ansatz der Margenmaximierung bei Support Vector Machines (SVM) den Kompromiss zwischen der Maximierung der Marge und der Minimierung von Klassifikationsfehlern ausbalanciert. Wie beeinflusst der Hyperparameter C dieses Gleichgewicht?

Aufgabe5: Clustering und Dimensionsreduktion

Erläutern Sie den Unterschied zwischen Clustering und Dimensionsreduktion im unüberwachten Lernen. Wie können Methoden wie k-Means und PCA miteinander kombiniert werden, um die explorative Datenanalyse zu verbessern? Geben Sie ein Beispiel, bei dem diese Techniken effektiv zusammenarbeiten.