Modulkonzept zu Overfitting und Underfitting

Förderkennzeichen:	01IS17073
Vorhabenbezeichnung:	Verbundprojekt: COSY-Entwicklung von sieben Praxis-Versuchen zum Thema Datenanalyse und Maschinelles Lernen an zwei Hochschulstandorten
Zugehörigkeit:	Modulkonzept Versuch 4 - Umweltdaten

Allgemeines:

Im Kontext des maschinellen Lernens beschreibt Overfitting den Effekt, dass die Hypothese zu genau an die Trainingsdaten angepasst wird, die Trainingsdaten also äuswendig gelernt"werden. Dies führt ab einem gewissen Zeitpunkt während des Trainings zur Verschlechterung der Hypothese auf den Testdaten. Deshalb sollte das Training rechzeitig abgebrochen werden bevor eine Überanpassung an die Trainingsdaten stattfindet. Um den richtigen Zeitpunkt für die Beendigung des Trainings zu finden wird die Hypothese während des Trainings auf einem zweiten, vom Trainingsset und Testset disjunkten, dritten Datensatz, dem sogenannten Validierungsset, getestet. Beginnt sich die Performance der Hypothese auf dem Validierungsset zu verschlechtern ist dies ein Zeichen dafür, dass eine Überanpassung stattfindet und das Training beendet werden sollte. Abbildung zeigt exemplarisch die Performance einer Hypothese während des Trainings auf dem Trainingsset und dem Validierungsset.

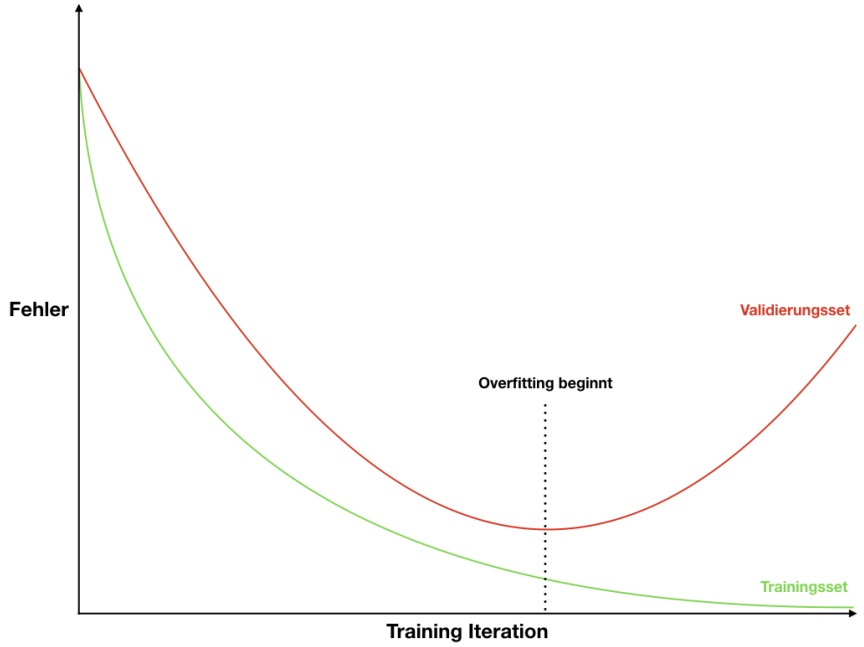

Abbildung 1: Fehler des Trainingssets und des Validierungssets während des Trainings

Zur Veranschaulichung dieses Effekts sollen die Studierenden mit der Matlab Toolbox Neuronale Netze jeweils mit und ohne Validierungsset trainieren und deren Qualität vergleichen. So sollen sie erkennen, dass Neuronale Netze die ohne Validierungsdaten trainiert wurden beim Test auf neuen Daten schlechter abschneiden als Netze bei denen Overfitting durch die Verwendung eines Validierungssets vermieden wurde.

Underfitting kann auftreten wenn die Dimensionalität im Verhältnis zur Anzahl der Samples zu hoch, beziehungsweise die Anzahl der Samples zu niedrig im Verhältnis zur Dimensionalität ist. In diesem Fall reichen die Samples nicht aus um eine hinreichend genaue Hypothese für das gesuchte System zu erstellen. Mit einer solchen Hypothese kann dann auch keine sinnvolle Prognose erstellt werden. Dieser Effekt kann mit einem Datensatz für den Heizenergieverbrauch demonstriert werden der zusätzlich auch die Temperaturen einzelner Räume betrachtet.

Bei der Durchführung einer Prognose mit einem Datensatz der die Temperaturen aller Einzelräume des Gebäudes betrachtet und eine mit dem gleichen Datensatz der jedoch nur eine gemittelte Temperatur für alle Räume betrachtet, stellt man fest, dass die zweite Prognose mit nur einem Mittelwert für die Raumtemperatur deutlich bessere Ergebnisse liefert.

Aufgaben:

- Overfitting:
  - Trainieren eines Neuronalen Netzes mit der Matlab Toolbox ohne Validierungsdaten
  - Trainieren eines Neuronalen Netzes mit der Matlab Toolbox mit Validierungsdaten
  - Vergleich der Qualität der beiden Neuronalen Netze
- Underfitting:
  - Durchführung einer Wärmeverbrauchsprognose mit linearer Regression auf einem Datensatz mit den Raumtemperaturen aller Einzelräume
  - Durchführung einer Wärmeverbrauchsprognose mit linearer Regression auf einem Datensatz mit einer gemittelten Raumtemperatur für die Einzelräume
  - Vergleich der Genauigkeit der beiden Prognosen

Ziele:

Die Studierenden sollen sich der Gefahr von Overfitting und Underfitting bewusst werden und verstehen wieso diese Effekte auftreten. Außerdem sollen die Studierenden Methoden und Ansatzpunkte (Validierungsset, Reduktion der Dimensionen, Verwendung von mehr Samples) kennen mit denen diese Effekte vermieden werden können.

Literatur:

- Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. (2012). Learning from data: a short course (S. 119-165). [United States]: AMLBook.com.