Fachgruppe Informatik
Prof. Dr. A. Ittner, C. Menzer
Wintersemester 2020/21
IF18/MI18

HOCHSCHULE MITTWEIDA
University of Applied Sciences

Big Data / Data Mining

Testmengen, Überanpassung

Testmengen

Bisher haben wir den Operator *Split Data* in RapidMiner einfach benutzt. Er teilt eine Menge von Beispielen prozentual in z.B. eine Trainings- und eine Testmenge.

Aufgabe 1

Lesen Sie die Seiten 231 bis 233 in Cleve und Lämmel (2014).

a) Wie sollte die Zerlegung von gegebenen Beispielen erfolgen und warum?

b) Was ist zu beachten, wenn verschiedene Data-Mining-Verfahren miteinander verglichen werden sollen?

c) Welches Problem kann bei zufälliger Auswahl der Testdaten auftreten?

d) Was ist unter Stratifikation zu verstehen?

e) Beschreiben Sie das Verfahren der Kreuzvalidierung. Stellen Sie das Prinzip grafisch dar. Welchen Vorteil bietet dieses Verfahren und wodurch entsteht er?

Überanpassung

Überanpassung (*overfitting*) tritt in der Statistik auf, wenn zu viele erklärende Variablen verwendet werden (Attribute/features). Beim Maschinellen Lernen kann zudem eine Überanpassung dadurch erfolgen, dass das Modell in den Trainingsschritten an die vorhandenen Trainingsdaten zu stark angepasst wird („auswendig lernen“). Solche Modelle ergeben oft keine guten Vorhersagen für neue Datensätze. Während die Fehlerrate auf den Trainingsdaten immer weiter sinkt, steigt sie hingegen auf den Testdaten wieder (vgl. Abbildung 1). Das Modell verliert somit an der Fähigkeit zu generalisieren.

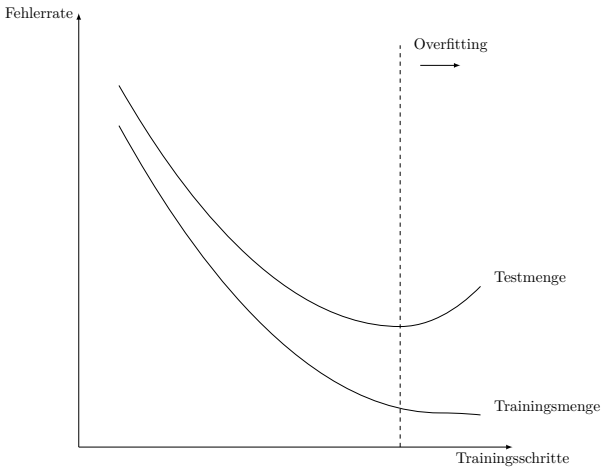

Abbildung 1: Vergleich der Fehlerrate bezogen auf die Anzahl der Trainingsschritte von Trainingsmenge und Testmenge

Aufgabe 2

a) Welche grundlegenden Maßnahmen können Sie gegen eine Überanpassung treffen?

b) Beurteilen Sie die Gefahr einer Überanpassung bei einem linearen Modell.

c) Wie können Sie einer Überanpassung bei Entscheidungsbäumen begegnen (vgl. Cleve und Lämmel 2014, S. 108 f.)? Wie unterscheiden sich die beiden grundlegenden Ansätze?

Aufgabe 3

Iris-Daten mit RapidMiner

Verwenden Sie den RapidMiner-Prozess aus dem Praktikum Entscheidungsbaum I und speichern Sie diesen unter einem neuen Namen im Repository.

a) Informieren Sie sich zum Operator *Cross Validation* (vgl. *Cross Validation - RapidMiner Documentation* 2020). Wenden Sie diesen Operator entsprechend an und loggen Sie die einzelnen Schritte während der Validierung mit. Messen Sie erneut die Genauigkeit.

b) Wenden Sie *Pruning* während des Entscheidungsbaumlernens an und vergleichen Sie die Genauigkeit.

Literatur

Cleve, Jürgen und Uwe Lämmel (2014). *Data mining*. Studium. München: De Gruyter Oldenbourg. 306 S. ISBN: 978-3-486-72034-1 978-3-486-71391-6.

Cross Validation - RapidMiner Documentation (2020). URL: https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html (besucht am 06. 01. 2020).