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

HOCHSCHULE MITTWEIDA
University of Applied Sciences

Big Data / Data Mining

k-Nearest-Neighbour

Aufgabe 1

Lesen Sie die Seiten 83 bis 89 in Cleve und Lämmel (2014).

Aufgabe 2

Klassifikation

a) Betrachten Sie Abbildung 1. Was hat entscheidenden Einfluss auf die Klassenzuordnung? Welchem Abstandsmaß entspricht die Darstellung der gepunkteten bzw. gestrichelten Kreise?

b) Geben Sie jeweils die Klasse für $k = 1$, $k = 3$ und $k = 5$ an. Verwenden Sie die Manhatten-Distanz.

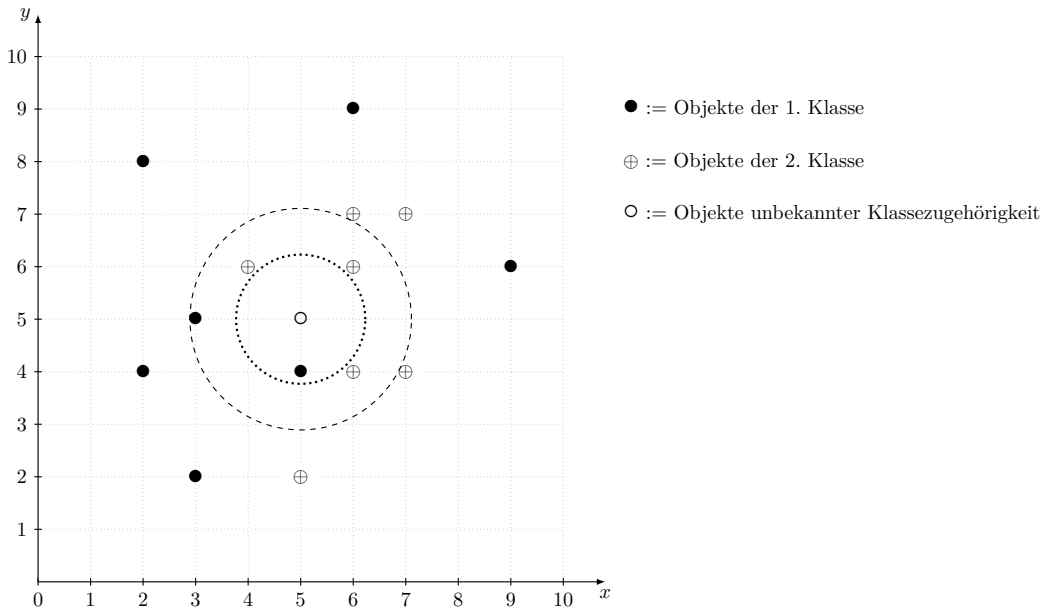

Abbildung 1: k-Nearest Neighbour im zweidimensionalen Raum

c) Betrachten Sie die Tabelle in Beispiel 5.1 aus ebd., S. 86 und übertragen Sie diese in ein Tabellenkalkulationsprogramm. Warum sollten Sie die Daten in dieser Form nicht verwenden? Korrigieren Sie die Daten mit geeigneten Funktionen Ihres Tabellenkalkulationsprogramms,

legen Sie dazu eine neue Spalte an und verwenden Sie die Formel aus ebd., S. 212. Ermitteln Sie jeweils den Abstand zum Zielattribut. Verwenden Sie die Euklidische Distanz.

d) *fakultativ:* Ermitteln Sie die Anzahl der Vorkommen jeder Klasse für alle k.

e) Wie können Sie mit nominalen Daten umgehen?

f) Wie können Sie den k-Nearest-Neighbour-Algorithmus verbessern?

Aufgabe 3

Vorhersage numerischer Werte

a) Worin unterscheidet sich der k-Nearest-Neighbour-Algorithmus vom diskreten Fall?

b) Modifizieren Sie die Tabelle aus Aufgabe 2 entsprechend, sodass das Zielattribut numerische Werte enthält. Sie können die Werte des Beispiels 5.3 aus ebd., S. 88 verwenden. Ermitteln Sie die Mittelwerte für alle k.

Aufgabe 4

Vergleich mit RapidMiner

a) Exportieren Sie geeignete Daten aus der Tabelle von Aufgabe 2.

b) Erstellen Sie einen RapidMiner-Prozess der die Daten importiert, normalisiert und ein Modell mittels k-NN erzeugt (Abbildung 2). Testen Sie verschiedene k und vergleichen Sie die Ergebnisse mit denen in der Tabelle aus Aufgabe 2.

Hinweis: Wenn Sie den neuen Datensatz ebenfalls in das *ExampleSet* importiert haben, müssen Sie diesen von den Trainingsdaten trennen. Sie können dafür den Operator *Split Data* verwenden.

Hinweis: RapidMiner liefert für jede Klasse einen Wert für die Konfidenz (Vertrauen). Wir werden diesen Begriff später genauer betrachten, allerdings unter den Gesichtspunkten der Assoziationsanalyse. Aber auch hier wird die bedingte Wahrscheinlichkeit berechnet: die relative Häufigkeit des Auftretens einer Klasse bei gegebener Klasse, vgl. ebd., S. 223 f. und Beispiel in LibreOffice Calc: Einkommen.ods.

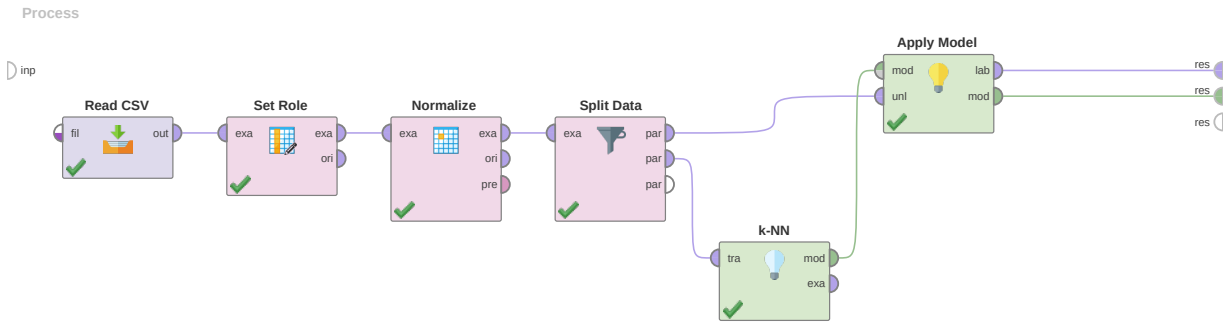

Abbildung 2: RapidMiner-Prozess für k-NN

Aufgabe 5

Körperfettanteil mit RapidMiner

a) Erstellen Sie ein Modell mittels k-NN mit einem Attribut. Ermitteln Sie den RMSE und vergleichen Sie diesen mit dem der Linearen Regression. Stellen Sie das Modell grafisch dar (Abbildung 3).

b) Optimieren Sie Ihr Modell, indem Sie die Attribute automatisiert auswählen lassen. Loggen Sie die einzelnen Werte für den RMSE, Anzahl sowie Namen der Attribute und wenden Sie das Modell auf Ihre Daten an.

c) Verbessern Sie Ihr Modell weiter, indem Sie automatisiert verschiedene k ausprobieren. Sehen Sie sich hierfür den Operator *Optimize Parameters (Grid)* an (*Optimize Parameters (Grid) - RapidMiner Documentation* 2019). Versuchen Sie die Ergebnisse einzuschätzen. Fallen Ihnen merkwürdige RMSE auf? Was könnte das Problem sein?

d) Verwenden Sie den Operator *Split Data* um die Testmenge von der Trainingsmenge zu trennen.

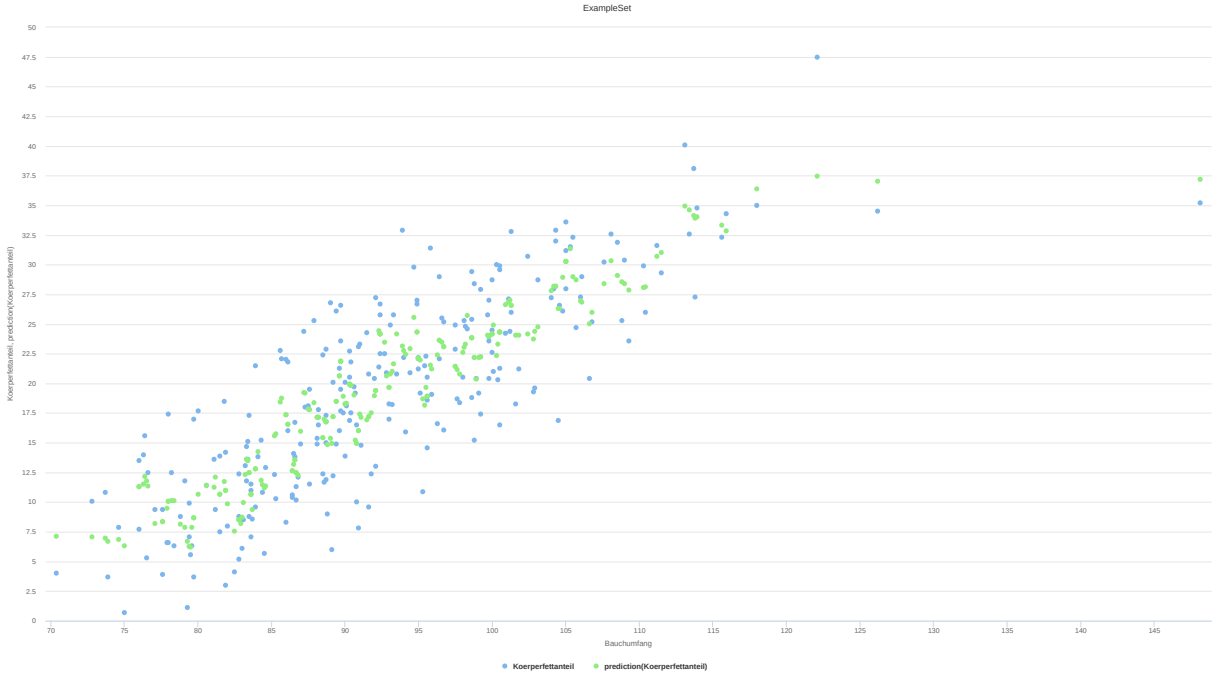

Abbildung 3: Grafische Darstellung der vorhergesagten Werte im Vergleich zu den gemessenen Werten

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.

Optimize Parameters (Grid) - RapidMiner Documentation (2019). URL: https://docs.rapidminer.com/latest/studio/operators/modeling/optimization/parameters/optimize_parameters_grid.html (besucht am 21. 10. 2019).