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

HOCHSCHULE MITTWEIDA
University of Applied Sciences

Big Data / Data Mining

Einführung

Aufgabe 1

Datentypen

a) Sie haben einige Datentypen kennengelernt, mit denen Data Mining konfrontiert sein kann. Ordnen Sie den Attributen den jeweiligen Datentyp zu:
   - Postleitzahl
   - ISBN
   - Kraftstoffverbrauch
   - Fahrzeugleistung
   - Hausnummer

b) Finden Sie für alle Datentypen mindestens zwei weitere Beispiele.

Aufgabe 2

Abstands- und Ähnlichkeitsmaße

Wir betrachten einige typische Distanzfunktionen:

Hamming-Distanz $\text{dist}_H(v, w) = \text{count}_i(v_i \neq w_i)$

Euklidische Distanz $\text{dist}_E(v, w) = \sqrt{\sum_i (v_i - w_i)^2}$

Manhatten-Distanz $\text{dist}_{Man}(v, w) = \sum_i |v_i - w_i|$

Maximum-Distanz $\text{dist}_{Max}(v, w) = \max(|v_i - w_i|)$

a) Informieren Sie sich zu diesen Distanzfunktionen in Cleve und Lämmel 2014, S. 43–46.

b) Berechnen Sie die Distanz zwischen den Punkten $(2, 3)$ und $(8, 7)$. Verwenden Sie alle 4 aufgeführten Distanzfunktionen (vgl. Abbildung 1).

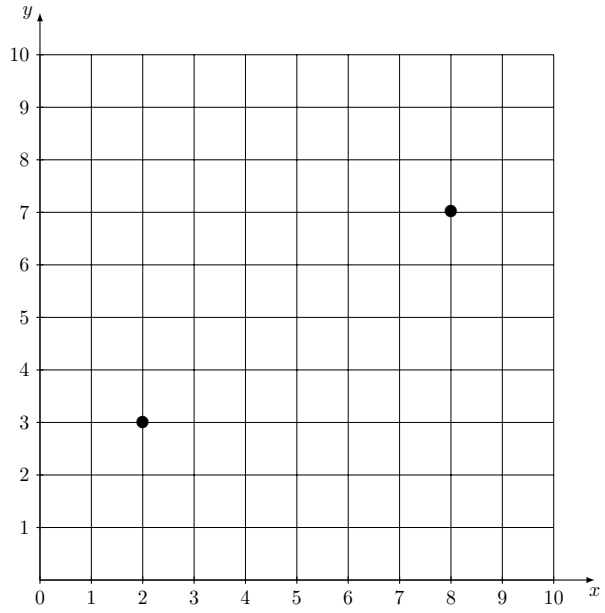

Abbildung 1: Distanzen im zweidimensionalen Raum

c) Berechnen Sie jeweils die Distanz zwischen den Punkten $(0, 1, 2)$, $(1, 5, 3)$ und $(4, -2, 3)$ für alle 4 Distanzfunktionen.

d) Begründen Sie, wieso die Hamming-Distanz unempfindlich gegen Ausreißer ist.

Aufgabe 3

Erste Schritte in RapidMiner

a) Registrieren Sie sich bei https://rapidminer.com/ und beziehen Sie eine Educational-Lizenz für RapidMiner Studio. Starten Sie RapidMiner Studio und synchronisieren Sie Ihre Lizenz im Lizenzmanager mit Ihrem Account.

Hinweis: Benutzen Sie während der Registrierung unbedingt Ihre Hochschul-E-Mail-Adresse. Sie können die Software ohne Installation starten. Ihr Dozent wird Ihnen die benötigten Dateien an einem geeigneten Ort bereitstellen. Auf privaten Geräten können Sie auch die Setups von der RapidMiner-Website verwenden und installieren.

b) Machen Sie sich mit dem Dashboard vertraut.

Lösungen

1. a)

- Postleitzahl: nominal
- ISBN: nominal
- Kraftstoffverbrauch: metrisch
- Fahrzeugleistung: metrisch
- Hausnummer: ordinal

2. b)

$$\begin{aligned} \text{dist}_H &= 2 \\ \text{dist}_E &\approx 7,2 \\ \text{dist}_{Man} &= 10 \\ \text{dist}_{Max} &= 6 \end{aligned}$$

2. c)

$(0,1,2)$ – $(1,5,3)$

$$\begin{aligned} \text{dist}_H &= 3 \\ \text{dist}_E &\approx 4,2 \\ \text{dist}_{Man} &= 6 \\ \text{dist}_{Max} &= 4 \end{aligned}$$

$(0,1,2)$ – $(4,-2,3)$

$$\begin{aligned} \text{dist}_H &= 3 \\ \text{dist}_E &\approx 5,1 \\ \text{dist}_{Man} &= 8 \\ \text{dist}_{Max} &= 4 \end{aligned}$$

$(1,5,3)$ – $(4,-2,3)$

$$\begin{aligned} \text{dist}_H &= 2 \\ \text{dist}_E &\approx 7,6 \\ \text{dist}_{Man} &= 10 \\ \text{dist}_{Max} &= 7 \end{aligned}$$

2. d) Die Werte der Attribute werden nicht berücksichtigt, lediglich ob sie sich vom Zielattribut unterscheiden.

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.