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 Nominal
- ISBN Nominal
- Kraftstoffverbrauch metrisch, kontinuierlich
- Fahrzeugleistung Metrisch, kontinuierlich
- Hausnummer metrisch, diskret

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

Nominal – Gesichtsausdrücke Wetter Ordinal – Farben Schulnoten metrisch diskret – Anzahl Fußballfans Anzahl Pixel kont. – Distanz Zeit

Aufgabe 2

Abstands- und Ähnlichkeitsmaße

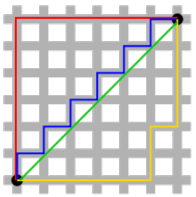
Wir betrachten einige typische Distanzfunktionen:

Hamming-Distanz $\text{dist}_H(v, w) = \text{count}_i(v_i \neq w_i)$ → vergleicht elementweise 0 1 1 0 0 1, 0 0 1 0 0 0 ⇒ 2

Euklidische Distanz $\text{dist}_E(v, w) = \sqrt{\sum_i (v_i - w_i)^2}$ → Abstand zweier Punkte im n-dim Raum

Manhatten-Distanz $\text{dist}_{Man}(v, w) = \sum_i |v_i - w_i|$ → Summe des Betrages aller Achsen/Dimensionen

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



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).

$u = (2,3)$ $v = (8,7)$

$\text{dist}_H(u,v) = 2$ $\text{dist}_{Man}(u,v) = 10$

$\text{dist}_E(u,v) = 7{,}2111$ $\text{dist}_{Max}(u,v) = 6$

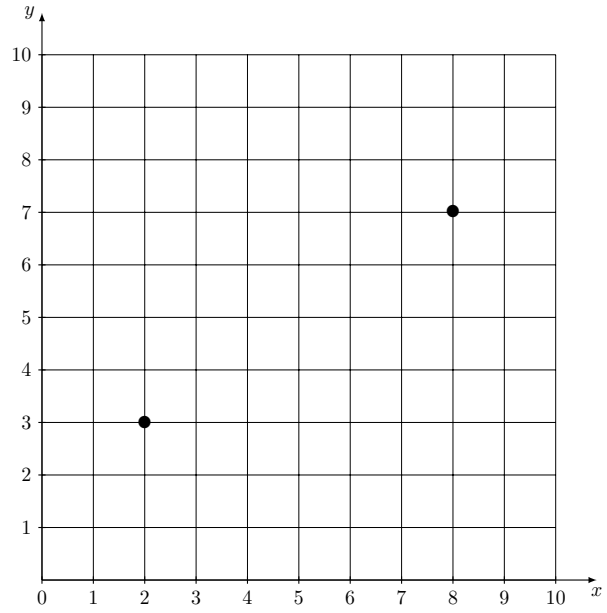


Abbildung 1: Distanzen im zweidimensionalen Raum

c)
$Dist_M(u,v) = 3$
$(u,w) = 3$
$(v,w) = 2$

$Dist_E(u,v) = \sqrt{17}$
$(u,w) = \sqrt{10}$
$(v,w) = \sqrt{58}$

$Dist_M(u,v) = 6$
$(u,w) = 8$
$(v,w) = 10$

$Dist_{Max}(u,v) = 4$
$(u,w) = 4$
$(v,w) = 7$

u = v = w =

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.

Weil Elemente nur paarweise auf Gleichheit geprüft werden

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.

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.