

Big Data Anwendungen

Aufgabenblatt 4

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- Aufgabenblatt 4 – Clusteringverfahren
 - Aufgabe 1 – Hierarchisches Clustering
 - Aufgabe 2 – kMeans Clustering
 - Aufgabe 3 – Verständnisfragen
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 1 (a) - Aufgabenstellung

- Gegeben seien folgende Daten:

Tag	Anzahl Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

- Nennen Sie alle Attribute für die mit Hilfe des Jaccard-Index keine Abstände ermittelt werden können und begründen Sie kurz. **(2 Punkte)**

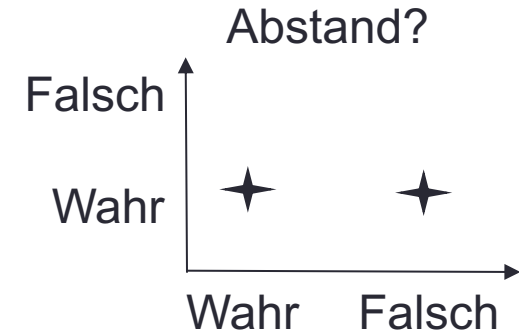
Wiederholung Vorlesung: Jaccard Abstand

- Idee
Zählen der Übereinstimmungen

- Grundlage
Jaccard Abstand: $d_J = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$

- Anwendung hier
$$d(p, q) = d(q, p) = \frac{n - \sum_{i=1}^n (p_i = q_i)}{n}$$
mit $p = (p_1, \dots, p_n), q = (q_1, \dots, q_n)$

- Beispiel
$$d(p, q) = d(q, p) = \frac{2-1}{2} = \frac{1}{2}$$



Aufgabe 1 (a) – Lösung (Jaccard Abstand)

- Sinnvolle Skalentype Jaccard Abstand
 - Kategorische Attribute
 - Überschaubar viele Ausprägungen
- Prinzipiell gilt
 - Jaccard-Index auf alle Skalentypen anwenden
 - Idee Prüfung auf Gleichheit für alle Attribute
- Hier
 - Sinnvolle Anwendung des Jaccard Abstand nur auf das Attribut “Regen”
 - Transformation für “Tag“, „Anzahl Sonnenstunden“ und „Temperatur“...
... ist nicht sinnvoll

Tag	Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

Aufgabe 1 (b) - Aufgabenstellung

- Gegeben seien folgende Daten:

Tag	Anzahl Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

- Nennen Sie zwei Verfahren zur Transformation von Daten auf ähnliche Wertebereiche. **(2 Punkte)**

Wiederholung Vorlesung: Transformationen

- Min-Max-Skalierung (Normierung auf $[0 \dots 1]$)
 - Formal: $x_i^t = \frac{x_i - \min_{j=1 \dots n} x_j}{\max_{j=1 \dots n} x_j - \min_{j=1 \dots n} x_j}$
 - Nachteil: Ausreißer beeinflussen Transformation stark
- Standardisierung (Normierung mit Standardabweichung und Mittelwert)
 - Formal: $x_i^t = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2}}$ mit $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$
 - Nachteil: Wertebereich nicht klar
- Rangtransformation (Normierung auf $[1 \dots n]$)
 - Abbilden des Wertes durch seine Position in aufsteigender Sortierung
 - Bei wiederholten auftreten identischer Werte
 - Vergabe mittleren Rangs oder
 - Vergabe des niedrigsten Rangs
 - Nachteil: Abstände zwischen Werten gehen verloren

Aufgabe 1 (b) – Lösung (Transformationen)

- Hier
 - Min-Max-Skalierung, Standardisierung und Rangtransformation für sinnvoll
 - Tag, Sonnenstunden und Temperatur lassen sich entsprechend abbilden
- Einzelne Attribute besitzen „Ausreißer“
 - Anzahl Sonnenstunden: 1x 10 sonst 2 bis 4
 - Temperatur: 1x 5 sonst -1 bis 2
- Damit bevorzugt anzuwenden
 - Standardisierung bzw. Rangtransformation
 - Präferenz für Standardisierung:
 - Ausreißer wirken plausibel (keine Datenfehler)
 - Abstände liefern zusätzliche Information

Tag	Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

Aufgabe 1 (c) - Aufgabenstellung

- Gegeben seien folgende Daten:

Tag	Anzahl Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

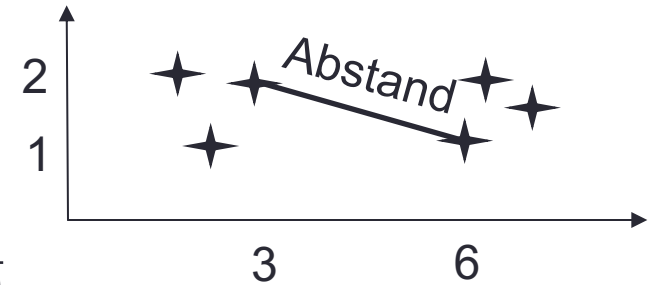
- Transformieren Sie die Attribute „Tag“, „Anzahl Sonnenstunden“ und „Temperatur“ so, dass Sie mit Hilfe der Euklidischen Distanz sinnvoll geclustert werden können. **(3 Punkte)**

Wiederholung Vorlesung: Abstandsmaße

- Euklidischer Abstand

- Länge direkter Verbindung zwischen Objekten
- Formal $d(p, q) = d(q, p) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$
mit $p = (p_1, \dots, p_n), q = (q_1, \dots, q_n)$
- Beispiel

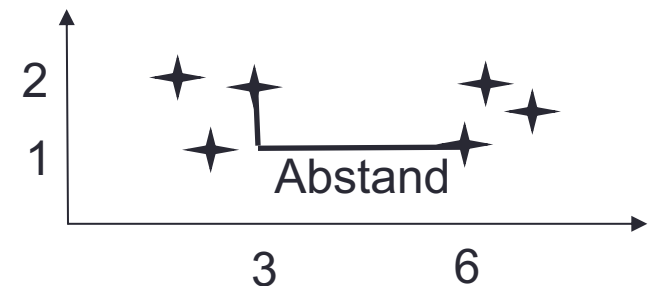
$$d(p, q) = d(q, p) = \sqrt{(2 - 1)^2 + (3 - 6)^2} = \sqrt{10}$$



- Manhattan Abstand

- Länge der Verbindung bei Lauf parallel zu Achsen
- Formal $d(p, q) = d(q, p) = \sum_{i=1}^n |q_i - p_i|$
mit $p = (p_1, \dots, p_n), q = (q_1, \dots, q_n)$
- Beispiel

$$d(p, q) = d(q, p) = |2 - 1| + |3 - 6| = 4$$



- Vorteil Manhattan Abstand:
Höhere Robustheit gegenüber Ausreißern

Aufgabe 1 (c) – Lösung (Standardisierung)

- Vorgehen

- Standardisierung

$$x_i^t = \frac{x_i - \bar{x}}{s}$$

mit $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$ und $s = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2}$

- Zunächst Ermittlung Mittelwerte ($\bar{x}$) und Standardabweichungen (s)

- Tag

- $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j = \frac{18+17+25+30}{4} = \frac{90}{4} = 22,50$

- $$s = \sqrt{\frac{1}{4-1} ((18 - 22,5)^2 + (17 - 22,5)^2 + (25 - 22,5)^2 + (30 - 22,5)^2)}$$
$$= \sqrt{\frac{1}{3} ((-4,5)^2 + (-5,5)^2 + (2,5)^2 + (7,5)^2)}$$
$$= \sqrt{\frac{1}{3} (20,25 + 30,25 + 6,25 + 56,25)} \approx \sqrt{37,67} \approx 6,14$$

Tag	Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

Aufgabe 1 (c) – Lösung (Standardisierung)

- Vorgehen

- Standardisierung

$$x_i^t = \frac{x_i - \bar{x}}{s}$$

$$\text{mit } \bar{x} = \frac{1}{n} \sum_{j=1}^n x_j \text{ und } s = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2}$$

- Zunächst Ermittlung Mittelwerte ($\bar{x}$) und Standardabweichungen (s)

Tag	Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

- Sonnenstunden

- $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j = \frac{10+2+4+3}{4} = \frac{19}{4} = 4,75$

- $$s = \sqrt{\frac{1}{4-1} ((10 - 4,75)^2 + (2 - 4,75)^2 + (4 - 4,75)^2 + (3 - 4,75)^2)}$$
$$= \sqrt{\frac{1}{3} ((5,25)^2 + (-2,25)^2 + (-0,75)^2 + (-1,75)^2)}$$
$$\approx \sqrt{\frac{1}{3} (27,56 + 7,56 + 0,56 + 3,06)} \approx \sqrt{12,92} \approx 3,59$$

Aufgabe 1 (c) – Lösung (Standardisierung)

- Vorgehen

- Standardisierung

$$x_i^t = \frac{x_i - \bar{x}}{s}$$

mit $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$ und $s = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2}$

- Zunächst Ermittlung Mittelwerte ($\bar{x}$) und Standardabweichungen (s)

Tag	Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

- Temperatur

- $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j = \frac{-1+5+1+2}{4} = \frac{7}{4} = 1,75$

- $s = \sqrt{\frac{1}{4-1} ((-1 - 1,75)^2 + (5 - 1,75)^2 + (1 - 1,75)^2 + (2 - 1,75)^2)}$
 $= \sqrt{\frac{1}{3} ((-2,75)^2 + (3,25)^2 + (-0,75)^2 + (0,25)^2)}$
 $\approx \sqrt{\frac{1}{3} (7,56 + 10,56 + 0,56 + 0,06)} \approx \sqrt{6,25} \approx 2,5$

Aufgabe 1 (c) – Lösung (Standardisierung)

- Bisher

- $\bar{x}$ und s für Attribute

Kennzahl	Tag	Sonnenstunden	Temperatur
$\bar{x}$	22,50	4,75	1,75
s	6,14	3,59	2,50

Tag	Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

- Jetzt Standardisierung: $x_i^t = \frac{x_i - \bar{x}}{s}$

- Ergebnis

Tag	Sonnenstunden	Temperatur
$\frac{18 - 22,50}{6,14} = -0,73$	$\frac{10 - 4,75}{3,59} = 1,46$	$\frac{-1 - 1,75}{2,50} = -1,10$
$\frac{17 - 22,50}{6,14} = -0,90$	$\frac{2 - 4,75}{3,59} = -0,77$	$\frac{5 - 1,75}{2,50} = 1,30$
$\frac{25 - 22,50}{6,14} = 0,41$	$\frac{4 - 4,75}{3,59} = -0,21$	$\frac{1 - 1,75}{2,50} = -0,30$
$\frac{30 - 22,50}{6,14} = 1,22$	$\frac{3 - 4,75}{3,59} = -0,49$	$\frac{2 - 1,75}{2,50} = 0,10$

Aufgabe 1 (d) - Aufgabenstellung

- Gegeben seien folgende Daten:

Tag	Anzahl Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

- Nutzen Sie hierarchisch agglomeratives Clustering zum Clustern der Daten aus (c). **(10 Punkte)**

Wiederholung Vorlesung: Hierarchisches Clustering

- Gegeben
Lerndatensatz mit g Beobachtungen mit n numerischen Attributen
(also, wie bisher: $g = (g_1, \dots, g_n)$)
- Gesucht
Partitionierung des Datensatz in Cluster (Keine Anzahl gegeben!)
- Initialisierung
 - Jede Beobachtung bildet eines von $k = |g|$ Clustern $z^i = (z_1^i, \dots, z_n^i)$
 - Ermittlung Distanz zwischen allen Clustern ($d(z^i, z^j)$) mit $i, j \in (1, \dots, k)$
- Reduktion der Clusterzentren
 - Identifikation der Cluster z^i, z^j mit dem geringsten Abstand...
... $((i, j) = \underset{i, j}{\operatorname{argmin}} d(z^i, z^j))$
 - Zusammenfassung der Cluster zu neuem Cluster z^l

Wiederholung Vorlesung: Hierarchisches Clustering

- Abbruchkriterium
Ist Anzahl der Cluster gleich 1 - Abbruch
- Aktualisierung der Daten
 - Ermittlung des Abstands zwischen neuem Cluster ...
... und alle bestehenden Clustern ($d(z^l, z^j)$) mit $j \in (1, \dots, k)$
 - Fortfahren mit Schritt „Reduktion der Clusterzentren“
- Anmerkungen
 - Anders als bei K-Means werden ein Mal getroffene...
... Gruppenzuordnungen nicht mehr geändert
 - Speicherbedarf bei großen Datensätzen groß verglichen mit K-Means
 - Bei K-Means: Distanz aller Knoten zu den k Clusterzentren
 - Hier: Distanz aller Cluster (initial aller Beobachtungen!) zueinander

Aufgabe 1 (d) – Lösung (Hierarchisches Clustering)

- Vorgehen
 - Initiales Clustering (jede Beobachtung)
 - Rekursiv
 - Ermittlung Abstand der Cluster
 - Kombination der zwei nächsten Cluster
- Initiales Clustering

Tag	Sonnenstunden	Temperatur
−0,73	1,46	−1,10
−0,90	−0,77	1,30
0,41	−0,21	−0,30
1,22	−0,49	0,10

Cluster	Tag	Sonnenstunden	Temperatur
A	−0,73	1,46	−1,10
B	−0,90	−0,77	1,30
C	0,41	−0,21	−0,30
D	1,22	−0,49	0,10

Aufgabe 1 (d) – Lösung (Hierarchisches Clustering)

- Vorgehen
 - Initiales Clustering (jede Beobachtung)
 - Rekursiv
 - Ermittlung Abstand der Cluster
 - Kombination der zwei nächsten Cluster

Cluster	Tag	Sonnenstunden	Temperatur
A	−0,73	1,46	−1,10
B	−0,90	−0,77	1,30
C	0,41	−0,21	−0,30
D	1,22	−0,49	0,10

Cluster	B	C	D
A	$\sqrt{\begin{aligned} &(-0,73 + 0,90)^2 \\ &+ (1,46 + 0,77)^2 \\ &+ (-1,10 - 1,30)^2 \end{aligned}} = 3,28$	$\sqrt{\begin{aligned} &(-0,73 - 0,41)^2 \\ &+ (1,46 + 0,21)^2 \\ &+ (-1,10 + 0,30)^2 \end{aligned}} = 2,17$	$\sqrt{\begin{aligned} &(-0,73 - 1,22)^2 \\ &+ (1,46 + 0,49)^2 \\ &+ (-1,10 - 0,10)^2 \end{aligned}} = 3,01$
B	-	$\sqrt{\begin{aligned} &(-0,90 - 0,41)^2 \\ &+ (-0,77 + 0,21)^2 \\ &+ (1,30 + 0,30)^2 \end{aligned}} = 2,14$	$\sqrt{\begin{aligned} &(-0,90 - 1,22)^2 \\ &+ (-0,77 + 0,49)^2 \\ &+ (1,30 - 0,10)^2 \end{aligned}} = 2,45$
C	-	Abstand von C und D am geringsten: Kombination	
			$\sqrt{\begin{aligned} &(0,41 - 1,22)^2 \\ &+ (-0,21 + 0,49)^2 \\ &+ (-0,30 - 0,10)^2 \end{aligned}} = 0,95$

Aufgabe 1 (d) – Lösung (Hierarchisches Clustering)

- Vorgehen
 - Initiales Clustering (jede Beobachtung)
 - Rekursiv
 - Ermittlung Abstand der Cluster
 - Kombination der zwei nächsten Cluster
- Kombination von C und D zu CD

Cluster	Tag	Sonnenstunden	Temperatur
A	−0,73	1,46	−1,10
B	−0,90	−0,77	1,30
C	0,41	−0,21	−0,30
D	1,22	−0,49	0,10

Cluster	Tag	Sonnenstunden	Temperatur
A	−0,73	1,46	−1,10
B	−0,90	−0,77	1,30
C	0,41	−0,21	−0,30
D	1,22	−0,49	0,10
CD	$\frac{0,41+1,22}{2} = 0,81$	$\frac{-0,21-0,49}{2} = -0,35$	$\frac{-0,30+0,10}{2} = -0,10$

Aufgabe 1 (d) – Lösung (Hierarchisches Clustering)

- Vorgehen
 - Initiales Clustering (jede Beobachtung)
 - Rekursiv
 - Ermittlung Abstand der Cluster
 - Kombination der zwei nächsten Cluster

Cluster	Tag	Sonnenstunden	Temperatur
A	−0,73	1,46	−1,10
B	−0,90	−0,77	1,30
CD	0,81	−0,35	−0,10

- Abstände der Cluster

Cluster	B	CD
A	$\sqrt{(-0,73 + 0,90)^2 + (1,46 + 0,77)^2 + (-1,10 - 1,30)^2} = 3,28$	$\sqrt{(-0,73 - 0,81)^2 + (1,46 + 0,35)^2 + (-1,10 + 0,10)^2} = 2,58$
B	-	$\sqrt{(-0,90 - 0,81)^2 + (-0,77 + 0,35)^2 + (1,30 + 0,10)^2} = 2,25$

Abstand von B und CD am geringsten:
Kombination

Aufgabe 1 (d) – Lösung (Hierarchisches Clustering)

- Vorgehen
 - Initiales Clustering (jede Beobachtung)
 - Rekursiv
 - Ermittlung Abstand der Cluster
 - Kombination der zwei nächsten Cluster

Cluster	Tag	Sonnenstunden	Temperatur
A	−0,73	1,46	−1,10
B	−0,90	−0,77	1,30
CD	0,81	−0,35	−0,10

- Kombination von C und D zu CD

Cluster	Tag	Sonnenstunden	Temperatur
A	−0,73	1,46	−1,10
B	−0,90	−0,77	1,30
C	0,41	−0,21	−0,30
D	1,22	−0,49	0,10
CD	$\frac{0,41+1,22}{2} = 0,81$	$\frac{-0,21-0,49}{2} = -0,35$	$\frac{-0,30+0,10}{2} = -0,10$
BCD	$\frac{-0,90+0,41+1,22}{3} = 0,24$	$\frac{-0,77-0,21-0,49}{3} = -0,49$	$\frac{1,30-0,30+0,10}{3} = 0,37$

Aufgabe 1 (d) – Lösung (Hierarchisches Clustering)

- Vorgehen
 - Initiales Clustering (jede Beobachtung)
 - Rekursiv
 - Ermittlung Abstand der Cluster
 - Kombination der zwei nächsten Cluster

Cluster	Tag	Sonnenstunden	Temperatur
A	−0,73	1,46	−1,10
BCD	0,24	−0,49	0,37

- Abstände der Cluster

Cluster	BCD
A	$\sqrt{(-0,73 - 0,24)^2 + (1,46 + 0,49)^2 + (-1,10 - 0,37)^2} = 2,63$

- Schritt für weitere Kombination irrelevant...
... für Erstellung des Dendrogramms aber nötig

Aufgabe 1 (e) - Aufgabenstellung

- Gegeben seien folgende Daten:

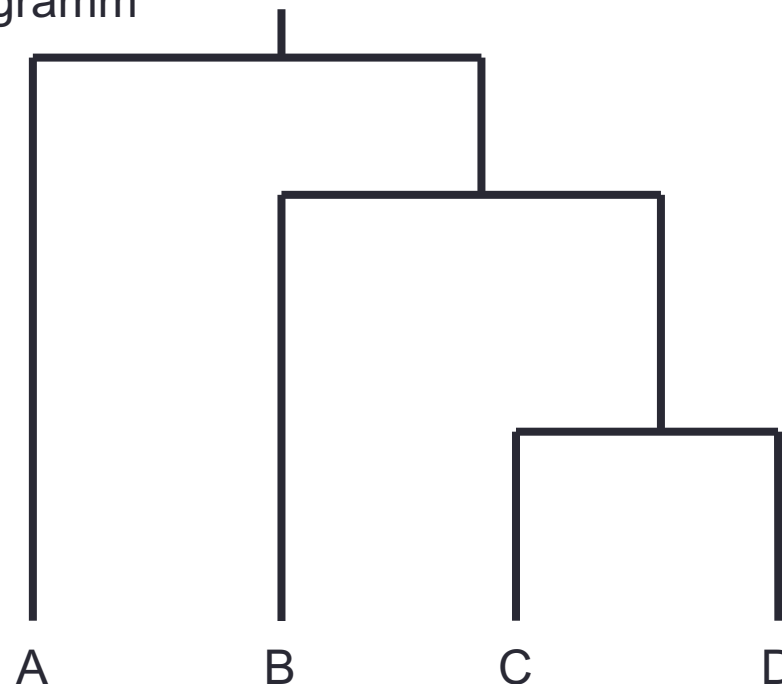
Tag	Anzahl Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

- Zeichnen Sie ein Dendrogramm für die Daten aus (d). **(3 Punkte)**

Aufgabe 1 (d) – Lösung (Dendrogramm)

- Bisher
 - Kombination der Cluster
 - Abstände der Cluster bei Kombination
- Jetzt: Ableitung Dendrogramm

Cluster	Abstände
C und D	0,95
B und CD	2,25
A und BCD	2,63



Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- **Aufgabenblatt 4 – Clusteringverfahren**
 - Aufgabe 1 – Hierarchisches Clustering
 - **Aufgabe 2 – kMeans Clustering**
 - Aufgabe 3 – Verständnisfragen
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 2 (a) - Aufgabenstellung

- Gegeben seien folgende Daten:

Höhe	Breite	Farbe	Preis
15	10	rot	50
20	15	gelb	50
17	15	rot	200
13	12	gelb	180
11	16	rot	110
15	12	gelb	120

- Geben Sie an welches Attribut Sie für ein kMeans Clustering ohne Transformation nicht nutzen können. Schlagen Sie eine Strategie vor, wie Sie dieses Attribut umwandeln können um es doch zu nutzen. **(3 Punkte)**

Wiederholung Vorlesung – kMeans Clustering

- Gegeben
Lerndatensatz mit g Beobachtungen mit n numerischen Attributen
(also, wie bisher: $g = (g_1, \dots, g_n)$)
- Gesucht
Partitionierung des Datensatz in k Cluster
- Initialisierung
Wähle zufällige k Cluster-Zentren mit $z^k = (z_1^k, \dots, z_n^k)$
- Zuordnung zu Clusterzentren
 - Ermittlung Euklidischer Distanz $(d(g, z^j) = \sqrt{\sum_{i=1}^n (g_i - z_i^j)^2}) \dots$
... für jeden Datenpunkt zu allen Cluster-Zentren ($j = (1, \dots, k)$)
 - Zuordnung des Datenpunkts g zu dem Clusterzentrum...
... mit geringstem Abstand ($j = \underset{j}{\operatorname{argmin}} d(g, z^j)$)

Wiederholung Vorlesung – kMeans Clustering

- Neuberechnung der Clusterzentren
 - Bilden der Beobachtungsmenge für jedes Clusterzentrum...
... $G_j = \{g | d(g, z^j) = \min d(g, z^j) \text{ mit } j \in (1, \dots, k)\}$
 - Bestimmung neuer Clusterzentren $z^k = (\sum_{g \in G_j} \frac{g_1}{|G_j|}, \dots, \sum_{g \in G_j} \frac{g_n}{|G_j|})$
 - Neues Clusterzentrum ist “Mittelwert”...
... über alle zugeordneten Beobachtungen
- Wiederholung der Schritte
„Zuordnung zu Clusterzentren“ und „Neuberechnung der Clusterzentren“...
... bis Clusternzentren stabil, d.h.
 - Keine Anpassung der Zuordnung der Beobachtungen oder
 - Verschiebung der Clusterzentren nur noch minimal

Aufgabe 2 (a) – Lösung (kMeans Clustering)

- Kritisch
 - kMeans Clustering nutzt Euklidische Distanz
 - Euklidische Distanz nur auf numerischen Attributen
- Problem
 - Attribut “Farbe” ist kategorisch
 - Anwendung hier nicht möglich
- Lösungsvorschlag
 - Umkodierung der Daten
 - Beispiel: 0 – gelb und 1 – rot

Höhe	Breite	Farbe	Preis
15	10	rot	50
20	15	gelb	50
17	15	rot	200
13	12	gelb	180
11	16	rot	110
15	12	gelb	120

Aufgabe 2 (b) - Aufgabenstellung

- Gegeben seien folgende Daten:

Höhe	Breite	Farbe	Preis
15	10	rot	50
20	15	gelb	50
17	15	rot	200
13	12	gelb	180
11	16	rot	110
15	12	gelb	120

- Sehen Sie sich die Daten an. Gehen Sie davon aus, dass Sie die Daten in 3 Cluster unterteilen wollen. Geben Sie an, welche Zuordnung der Datenpunkte zu Clustern Sie erwarten, wenn Sie die Euklidische Distanz als Distanzmaß wählen. Erörtern Sie kurz, wie sie zu einem weniger vorhersehbaren Ergebnis gelangen können.

(3 Punkte)

Aufgabe 2 (b) – Lösung (kMeans Clustering)

- Hier relevant
 - Euklidische Distanz auf numerischen Attributen
 - Attribute gehen identisch in Euklidische Distanz ein
 - Je größer der Wertebereich eines Attributs...
... umso größer die Bedeutung des Attributs
- Konsequenz
 - Wertebereich von Preis deutlich größer als andere Wertebereiche
 - Erwartetes Clustering
 - Cluster 1: 1. und 2. Beobachtung
 - Cluster 2: 3. und 4. Beobachtung
 - Cluster 3: 5. und 6. Beobachtung
- Lösungsmöglichkeit
 - Transformation der Datenpunkt auf ähnliche Intervalle
 - Konkret: Standardisierung, Min-Max-Skalierung oder Rangtransformation

Höhe	Breite	Farbe	Preis
15	10	rot	50
20	15	gelb	50
17	15	rot	200
13	12	gelb	180
11	16	rot	110
15	12	gelb	120

Aufgabe 2 (c) - Aufgabenstellung

- Gegeben seien folgende Daten:

Höhe	Breite	Farbe	Preis
15	10	rot	50
20	15	gelb	50
17	15	rot	200
13	12	gelb	180
11	16	rot	110
15	12	gelb	120

- Wenden Sie den kMeans Algorithmus an um die Daten auf Basis der 3 numerischen Attribute in 3 Cluster zu zerlegen. Nutzen Sie dabei die Euklidische Distanz und lösen Sie das Problem rechnerisch. **(12 Punkte)**

Aufgabe 2 (c) – Lösung (kMeans Clustering)

- Vorgehen
 - Initiale Clusterzentren
 - Rekursiv
 - Ermittlung Abstand zu den Clustern
 - Zuordnung zu den Clustern
 - Ermittlung neuer Clusterzentren

Höhe	Breite	Farbe	Preis
15	10	rot	50
20	15	gelb	50
17	15	rot	200
13	12	gelb	180
11	16	rot	110
15	12	gelb	120

- Initiale Clusterzentren

	Höhe	Breite	Preis
Zentrum 1	15	10	50
Zentrum 2	17	15	200
Zentrum 3	11	16	110

- Anmerkung
 - Wahl der Clusterzentren auf Basis der Ergebnisse in (a) und (b)
 - Clusterzentren sollten schnell zum Ziel führen

Aufgabe 2 (c) – Lösung (kMeans Clustering)

- Vorgehen
 - Initiale Clusterzentren
 - Rekursiv
 - Ermittlung Abstand zu den Clustern
 - Zuordnung zu den Clustern
 - Ermittlung neuer Clusterzentren
- Ermittlung Abstand zu den Clustern

	Höhe	Breite	Preis
Beobachtung	15	10	50
	20	15	50
	17	15	200
	13	12	180
	11	16	110
	15	12	120
Z. 1	15	10	50
Z. 2	17	15	200
Z. 3	11	16	110

ID	Distanz zu Zentrum 1	Distanz zu Zentrum 2	Distanz zu Zentrum 3
1	0	$\sqrt{2^2 + 5^2 + 150^2} = 150,10$	$\sqrt{4^2 + 6^2 + 60^2} = 62,39$
2	$\sqrt{5^2 + 5^2 + 0^2} = 7,07$	$\sqrt{3^2 + 0^2 + 150^2} = 150,03$	$\sqrt{9^2 + 1^2 + 60^2} = 60,68$
3	$\sqrt{2^2 + 5^2 + 150^2} = 150,10$	0	$\sqrt{6^2 + 1^2 + 60^2} = 90,21$
4	$\sqrt{2^2 + 2^2 + 130^2} = 130,03$	$\sqrt{4^2 + 3^2 + 20^2} = 20,62$	$\sqrt{2^2 + 4^2 + 70^2} = 70,14$
5	$\sqrt{4^2 + 6^2 + 60^2} = 60,43$	$\sqrt{6^2 + 1^2 + 90^2} = 90,21$	0
6	$\sqrt{2^2 + 0^2 + 70^2} = 70,03$	$\sqrt{2^2 + 3^2 + 80^2} = 80,08$	$\sqrt{4^2 + 4^2 + 10^2} = 11,49$

Aufgabe 2 (c) – Lösung (kMeans Clustering)

- Vorgehen
 - Initiale Clusterzentren
 - Rekursiv
 - Ermittlung Abstand zu den Clustern
 - Zuordnung zu den Clustern
 - Ermittlung neuer Clusterzentren
- Zuordnung zu den Clustern

	Höhe	Breite	Preis
Beobachtung	15	10	50
	20	15	50
	17	15	200
	13	12	180
	11	16	110
	15	12	120
Z. 1	15	10	50
Z. 2	17	15	200
Z. 3	11	16	110

ID	Distanz zu Z. 1	Distanz zu Z. 2	Distanz zu Z. 3	Zuordnung zu
1	0,00	150,10	62,39	Z. 1
2	7,07	150,03	60,68	Z. 1
3	150,10	0,00	90,21	Z. 2
4	130,03	20,62	70,14	Z. 2
5	60,43	90,21	0,00	Z. 3
6	70,03	80,08	11,49	Z. 3

Aufgabe 2 (c) – Lösung (kMeans Clustering)

- Vorgehen
 - Initiale Clusterzentren
 - Rekursiv
 - Ermittlung Abstand zu den Clustern
 - Zuordnung zu den Clustern
 - Ermittlung neuer Clusterzentren
- Ermittlung neuer Clusterzentren

	Höhe	Breite	Preis
Beobachtung	15	10	50
	20	15	50
	17	15	200
	13	12	180
	11	16	110
	15	12	120
Z. 1	1. und 2. Beobachtung		
Z. 2	3. und 4. Beobachtung		
Z. 3	5. und 6. Beobachtung		

	Höhe	Breite	Preis
Zentrum 1	$(15 + 20)/2 = 17,5$	$(10 + 15)/2 = 12,5$	$(50 + 50)/2 = 50$
Zentrum 2	$(17 + 13)/2 = 15,0$	$(15 + 12)/2 = 13,5$	$(200 + 180)/2 = 190$
Zentrum 3	$(11 + 15)/2 = 13,0$	$(16 + 12)/2 = 14,0$	$(110 + 120)/2 = 115$

Aufgabe 2 (c) – Lösung (kMeans Clustering)

- Vorgehen
 - Initiale Clusterzentren
 - Rekursiv
 - Ermittlung Abstand zu den Clustern
 - Zuordnung zu den Clustern
 - Ermittlung neuer Clusterzentren
- Ermittlung Abstand zu den Clustern

	Höhe	Breite	Preis
Beobachtung	15	10	50
	20	15	50
	17	15	200
	13	12	180
	11	16	110
	15	12	120
Z. 1	17,5	12,5	50
Z. 2	15,0	13,5	190
Z. 3	13,0	14,0	115

ID	Distanz zu Zentrum 1	Distanz zu Zentrum 2	Distanz zu Zentrum 3
1	$\sqrt{2,5^2 + 2,5^2 + 0^2} = 3,54$	$\sqrt{0^2 + 3,5^2 + 40^2} = 140,04$	$\sqrt{2^2 + 4^2 + 65^2} = 65,15$
2	$\sqrt{2,5^2 + 2,5^2 + 0^2} = 3,54$	$\sqrt{5^2 + 1,5^2 + 40^2} = 140,10$	$\sqrt{7^2 + 1^2 + 65^2} = 65,38$
3	$\sqrt{0,5^2 + 2,5^2 + 150^2} = 150$	$\sqrt{2^2 + 1,5^2 + 10^2} = 10,31$	$\sqrt{4^2 + 1^2 + 85^2} = 85,10$
4	$\sqrt{4,5^2 + 0,5^2 + 130^2} = 130$	$\sqrt{2^2 + 1,5^2 + 10^2} = 10,31$	$\sqrt{0^2 + 2^2 + 65^2} = 65,03$
5	$\sqrt{6,5^2 + 3,5^2 + 60^2} = 60,45$	$\sqrt{4^2 + 2,5^2 + 80^2} = 80,14$	$\sqrt{2^2 + 2^2 + 5^2} = 5,74$
6	$\sqrt{2,5^2 + 0,5^2 + 70^2} = 70,05$	$\sqrt{0^2 + 1,5^2 + 70^2} = 70,02$	$\sqrt{2^2 + 2^2 + 5^2} = 5,74$

Aufgabe 2 (c) – Lösung (kMeans Clustering)

- Vorgehen
 - Initiale Clusterzentren
 - Rekursiv
 - Ermittlung Abstand zu den Clustern
 - Zuordnung zu den Clustern
 - Ermittlung neuer Clusterzentren
- Zuordnung zu den Clustern

	Höhe	Breite	Preis
Beobachtung	15	10	50
	20	15	50
	17	15	200
	13	12	180
	11	16	110
	15	12	120
Z. 1	15	10	50
Z. 2	17	15	200
Z. 3	11	16	110

ID	Distanz zu Z. 1	Distanz zu Z. 2	Distanz zu Z. 3	Zuordnung zu
1	3,54	140,04	65,15	Z. 1
2	3,54	140,10	65,38	Z. 1
3	150,00	10,31	85,10	Z. 2
4	130,00	10,31	65,03	Z. 2
5	60,45	80,14	5,74	Z. 3
6	70,05	70,02	5,74	Z. 3

- Keine weiteren Schritte nötig, da Clusterzentren unverändert

Agenda

- Aufgabenblatt 1 – Deskriptive Methoden zur Datenexploration
- Aufgabenblatt 2 – Klassifikation
- Aufgabenblatt 3 – Recommender Systems
- **Aufgabenblatt 4 – Clusteringverfahren**
 - Aufgabe 1 – Hierarchisches Clustering
 - Aufgabe 2 – kMeans Clustering
 - **Aufgabe 3 – Verständnisfragen**
- Aufgabenblatt 5 – Stream Mining
- Aufgabenblatt 6 – Social Network Analysis

Aufgabe 3 (a) - Aufgabenstellung

- Erläutern Sie, wie k Nearest Neighbor Klassifikation funktioniert. Diskutieren Sie dann Vor- und Nachteile des Verfahrens gegenüber anderen Klassifikationsalgorithmen. **(6 Punkte)**

Wiederholung: kNearest Neighbor Klassifikation

- Algorithmus: K-Nearest-Neighbor Klassifikation
- Gegeben:
 - Trainingsdatensatz T
- Vorgehen zur Klassifikation einer Beobachtung g_i :
 - Menge S : k ($k \bmod 2 = 1$) nächsten Nachbarn von g_i in T
 - Klassifikation von g_i durch Klasse mit meisten Elementen in S
- Anmerkungen:
 - „Nähe“ über Distanzmaß (z.B. euklidische oder Manhattan Distanz)
 - Kein „Lernen“ wie bei anderen Klassifikationsalgorithmen nötig
 - Klassifikation mit großen Trainingsdatensätzen T aufwendig
⇒ Einsatz repräsentativer Stichproben von T
- Anwendung auf Regressionsprobleme
Beispielsweise durch Rückgabe des Mittelwerts der k-Nearest-Neighbors

Aufgabe 3 (a) – Lösung (kNearest Neighbor)

- Vorteile
 - Kein Lernen notwendig
 - Anwendung auf Regressionsprobleme möglich
 - Funktioniert gut bei numerischen Attributen
- Nachteile
 - Aufwendig bei großen Trainingsdaten
 - Trainingsdaten müssen repräsentativ sein
 - Transformation der Attribute nötig

Aufgabe 3 (b) - Aufgabenstellung

- Erläutern Sie kurz was der Unterschied zwischen supervised und unsupervised learning Verfahren ist. **(2 Punkte)**

Aufgabe 3 (b) – Lösung (Unsupervised Learning)

- Supervised learning
 - Vorherzusagende Eigenschaft bekannt
 - Typische Beispiele: Klassifikation, Regression
- Unsupervised learning
 - Vorherzusagende Eigenschaft nicht bekannt
 - Typisches Beispiel: Clustering

Aufgabe 3 (c) - Aufgabenstellung

- Erläutern Sie warum die Rangskalierung von Attributen für die Anwendung von Clusteringverfahren hilfreich sein kann und gehen Sie auf Nachteile des Ansatzes ein. **(2 Punkte)**

Aufgabe 3 (c) – Lösung (Rangskalierung)

- Vorteile
 - Abstände zwischen Attributen werden eliminiert
 - Vorhersagbarer Wertebereich
 - Eliminierung von Ausreißern
 - Anonymisierung der Daten
- Nachteile
 - Abstände zwischen Attributen werden eliminiert
 - Ausreißer werden nicht berücksichtigt
 - Daten müssen sortiert werden
 - Kaum übertragbar auf neue Daten