Big Data Anwendungen

Clusteringverfahren

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems

- Clusteringverfahren
  - Prinzip und Distanzmaße
  - Übersicht Verfahren
  - K-Means
  - Hierarchische Verfahren
  - Anwendungsbeispiel

- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

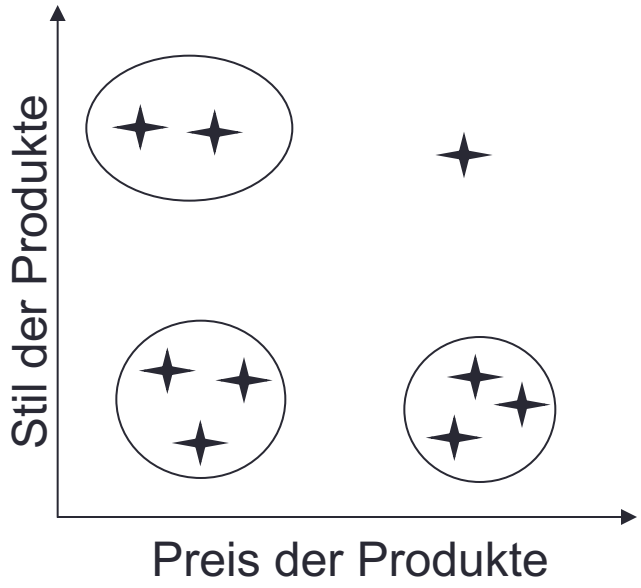
Grundidee

- Ziel
  - Ermittlung der Unterschiedlichkeit von Objekten
  - Bilden von Gruppen (Clustern) ähnlicher Objekte
  - Identifikation der Anzahl von Gruppen (optional)

- Bewertung der Cluster
  - Hohe Ähnlichkeit innerhalb eines Clusters
  - Große Verschiedenheit zwischen Clustern
  - Kenngröße: „Abstandsmaß“

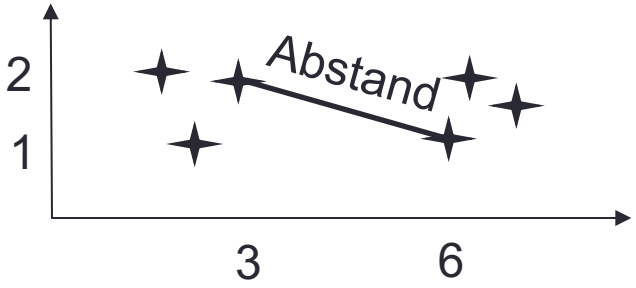
- Clustering ist „Unsupervised Learning“
  - Bisher primär Supervised Learning: Zielattribut ist beim Lernen bekannt
  - Hier ist Zuordnung zu Clustern unklar

- Herausforderung: Modellbewertung!

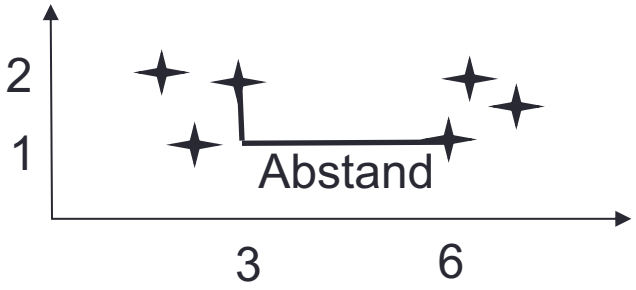


Abstandsmaße – kontinuierliche Attribute

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



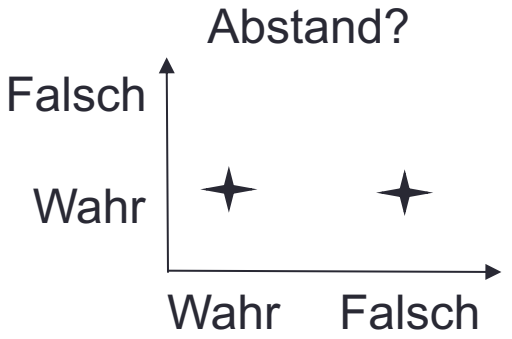
- Manhattan Abstand
  - Länge der Verbindung bei Lauf parallel zu Achsen
  - Formal $d(p,q) = d(q,p) = \sum_{i=1}^{n} |q_i - p_i|$ mit $p = (p_1, \dots, p_n), q = (q_1, \dots, q_n)$
  - Beispiel
    $d(p,q) = d(q,p) = |2-1| + |3-6| = 4$



- Vorteil Manhattan Abstand:
  Höhere Robustheit gegenüber Ausreißern

Abstandsmaße – kategorische / binäre Attribute

- Nachteile: Euklidischer und Manhattan Abstand
  - Nicht sinnvoll für kategorische Attribute (Bsp.: Geschlecht, Browsereigenschaften)
  - Identifikation von Clustern hier aber oft sinnvoll
  - „Einfache“ Lösung: Übertragung der Eigenschaften in Wahrscheinlichkeiten

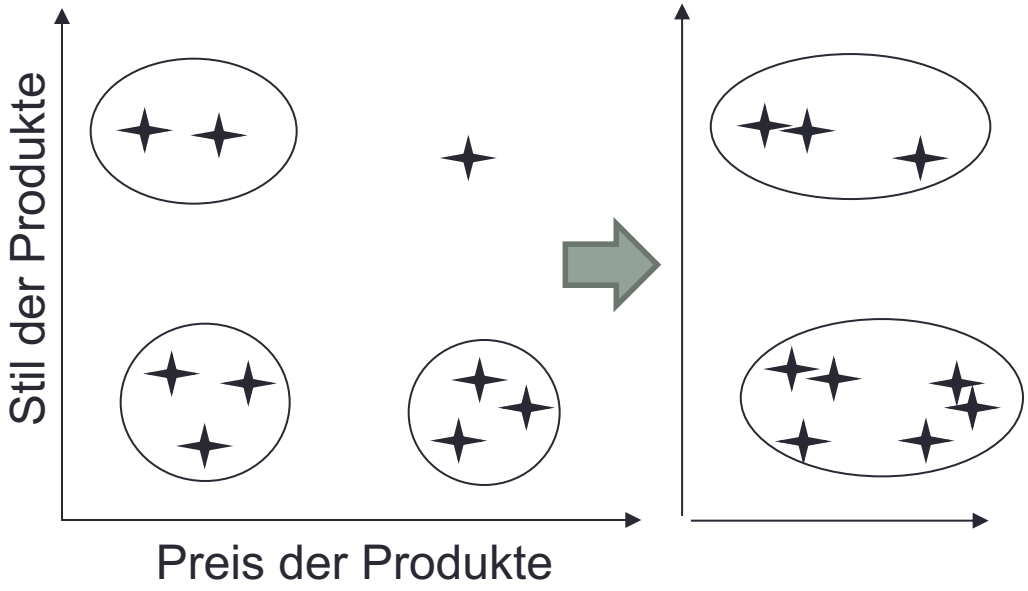


- Jaccard Abstand (weitere Lösung)
  - Zählen der Übereinstimmungen
  - Grundlage: Jaccard Abstand: $\mathrm{d_J} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$
  - Anwendung hier: $d(p,q) = d(q,p) = \frac{n - \sum_{i=1}^{n}(p_i = q_i)}{n}$ mit $p = (p_1, \ldots, p_n), q = (q_1, \ldots, q_n)$
  - Beispiel $d(p,q) = d(q,p) = \frac{2-1}{2} = \frac{1}{2}$

Transformation von Clusteringattributen I

- Problem:
  Skalierung metrischer Ab-...
  ... stände beeinflusst Clustering

- Lösung: Transformation
  Sicherstellen, dass Werte-...
  ...bereich aller Attribute gleich



- Ansätze Skalierung
  - Normierung auf $[0 \ldots 1]$
    - Formal: $x_i^t = \frac{x_i - \min_{j=1\ldots n} x_j}{\max_{j=1\ldots n} x_j - \min_{j=1\ldots n} x_j}$
    - Nachteil: Ausreißer beeinflussen Transformation stark
  - 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

Transformation von Clusteringattributen II

- Ansatz Rangtransformation
  - Aufsteigende Sortierung aller Werte
  - Vergeben einer aufsteigenden Zahl für jeden Wert
  - Bei wiederholten auftreten identischer Werte
    - Vergabe mittleren Rangs oder
    - Vergabe des niedrigsten Rangs
  - Beispiel

	345	23452	23452	234554	234556
Mittlerer Rang	1	2,5	2,5	4	5
Niedrigster Rang	1	2	2	4	5

  - Nachteil
    Abstände zwischen Werten gehen verloren

- Ansatz Ausreißerbehandlung
  - Eliminierung von Ausreißern

Agenda

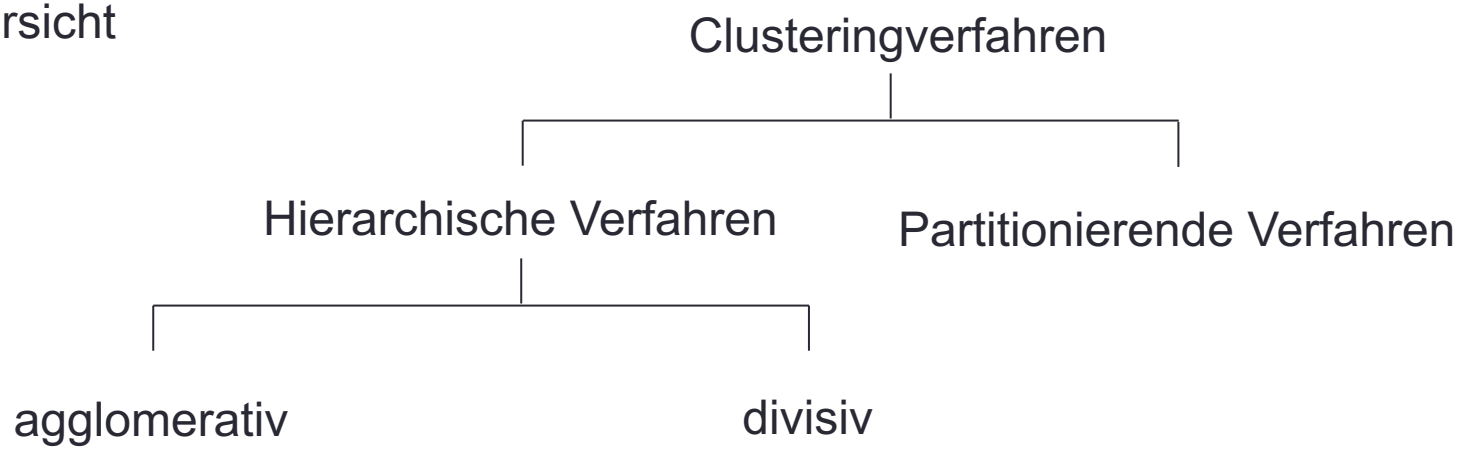
- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems

- Clusteringverfahren
  - Prinzip und Distanzmaße
  - Übersicht Verfahren
  - K-Means
  - Hierarchische Verfahren
  - Anwendungsbeispiel

- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Kategorisierung

- Übersicht



- Partitionierende Verfahren (z.B. k-Means, k-Medoids)
  - Initial zufällige Wahl von Clusterzentren
  - Iterative Anpassung bis „Gütekriterien“ erreicht
- Hierarchische Verfahren
  - Divisive Verfahren
    - Ausgehend von einem Cluster
    - Wiederholte Splits bis Gütekriterien erreicht
  - Agglomerativer Verfahren
    - Ausgehend von ein-elementigen Clustern
    - Wiederholte Zusammenführung ähnlicher Cluster

Zuordnungsprinzipien zu Clustern

- Deterministische Zuordnung
  Beobachtungen werden mit Wahrscheinlichkeit 1 zugordnet
  - Einem Cluster („nicht-überlappende Zuordnung“)
  - Mehreren Clustern („überlappende Zuordnung“)

- Probabilistische Zuordnung
  Beobachtungen werden mit Wahrscheinlichkeit zwischen 0 und 1...
  ... einem oder mehreren Clustern zugeordnet

- Possibilistische Zuordnung
  Beobachtungen werden mit Wahrscheinlichkeit zwischen 0 und 1...
  ... jedem verfügbaren Cluster zugeordnet

- Nutzen possibilistischer Zuordnung
  - Kunden wandern zwischen Clustern
    ⇒ Kunden auf Grenze zwischen zwei Clustern gesondert ansprechen
  - Aktueller Trend: Individualisierung
    ⇒ Kundenindividuelle Kombination von Werbemitteln über alle Cluster

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems

- Clusteringverfahren
  - Prinzip und Distanzmaße
  - Übersicht Verfahren
  - K-Means
  - Hierarchische Verfahren
  - Anwendungsbeispiel

- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Vorgehen I

- 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 ($d(g, z^j) = \sqrt{\sum_{i=1}^{n} (g_i - z_i^j)^2}$)...
    ... für jeden Datenpunkt zu allen Cluster-Zentren ($j = (1, \dots k)$)
  - Zuordnung des Datenpunkts g zu dem Clusterzentrum...
    ... mit geringstem Abstand ($j = \underset{j}{\operatorname{argmin}}\, d(g, z^j)$)

Vorgehen II

- 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

Anmerkungen

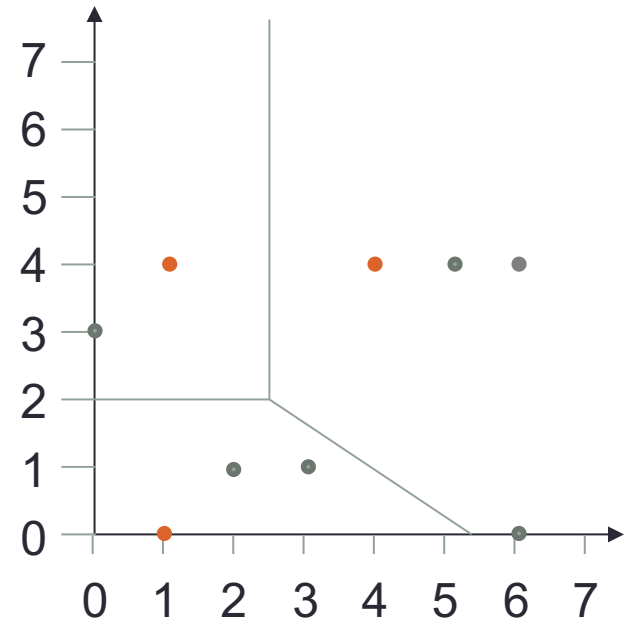
- Algorithmus ist empfindlich bzgl. initialer Clusterzentren
  - Findet verschiedene Clusterzuordnungen bei wiederholter Ausführung
  - Maßnahme: Wiederholte Ausführung mit unterschiedlichen Zentren
- Anzahl der Cluster k muss angegeben werden
  - Keine Möglichkeit zu Abschätzung welche Clusterzahl sinnvoll
  - Maßnahme: Bestimmung der Anzahl der Cluster separater Prüfung
- Anfällig für Outlier und Rauschen, da Verwendung der Mittelwerte
  - Algorithmus lässt jedoch Anwendung andere Distanzmaße zu
  - Maßnahme
    - Austausch des Distanzmaßes durch Manhattan-Distanz
    - Bereinigung des Datensatzes um Outlier vor Anwendung
- Keine Unterstützung kategorischer Attribute
  - Angewandte Distanzmaße erlauben nur numerische Eingaben
  - Maßnahme: Anwendung anderer Distanzmaße (z.B. Jaccard)

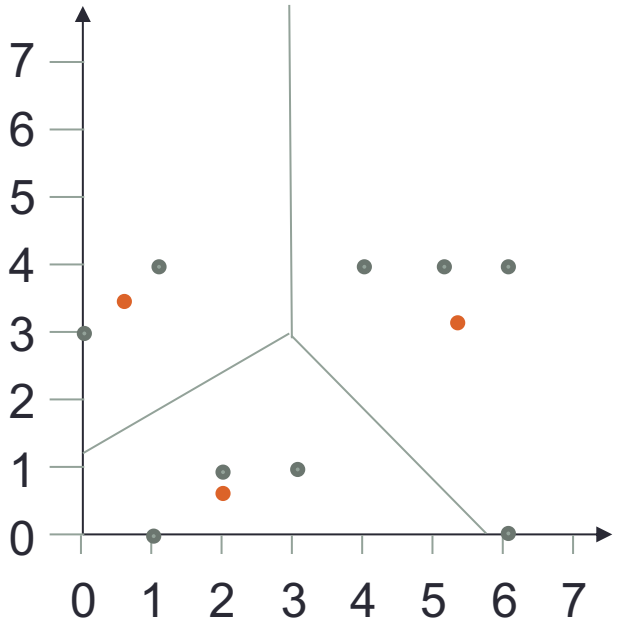
Beispiel – Graphische Lösung (Voronoi-Diagramm) I

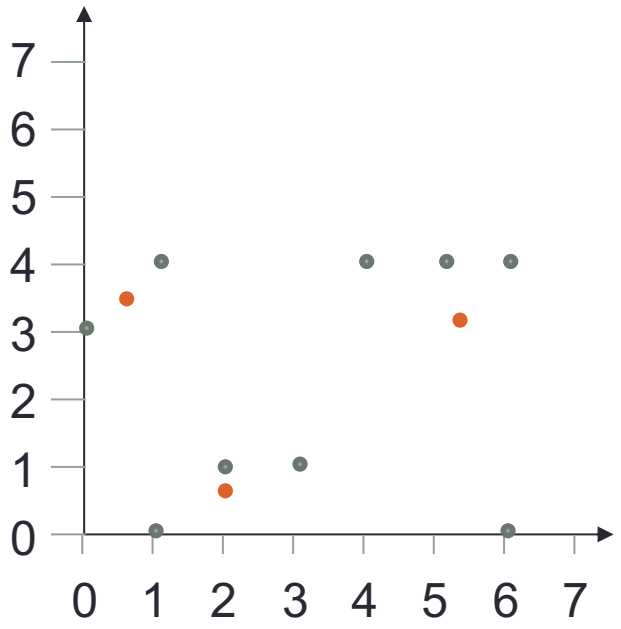
- 9 Beobachtungen (A bis I) mit zwei numerischen Attributen (x, y)

	A	B	C	D	E	F	G	H	I
x	0	1	1	2	3	4	5	6	6
y	3	4	0	1	1	4	4	4	0

- Anwendung von K-Means mit „günstigen“ Initialen Clusterzentren





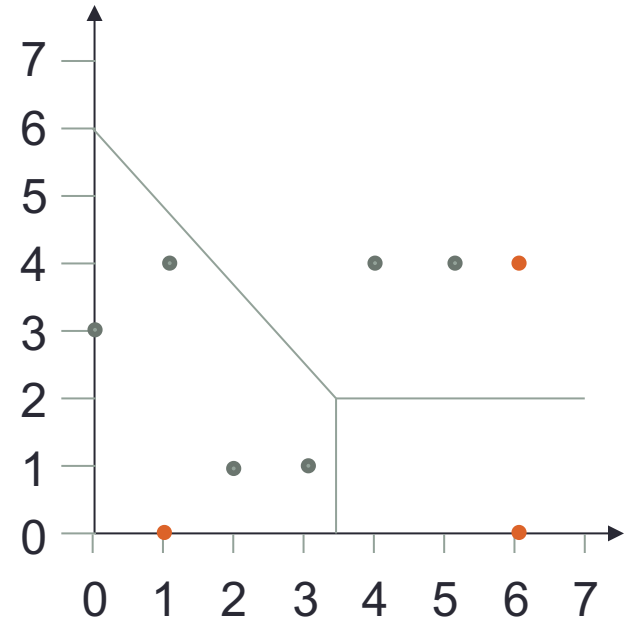


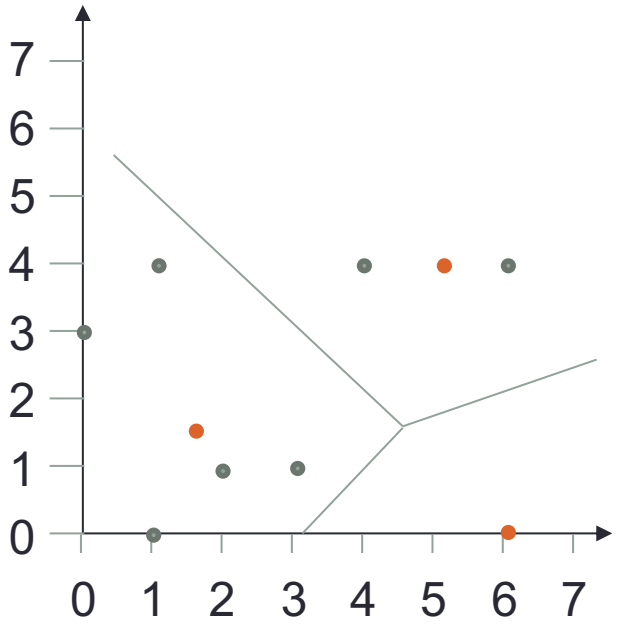
Beispiel – Graphische Lösung II

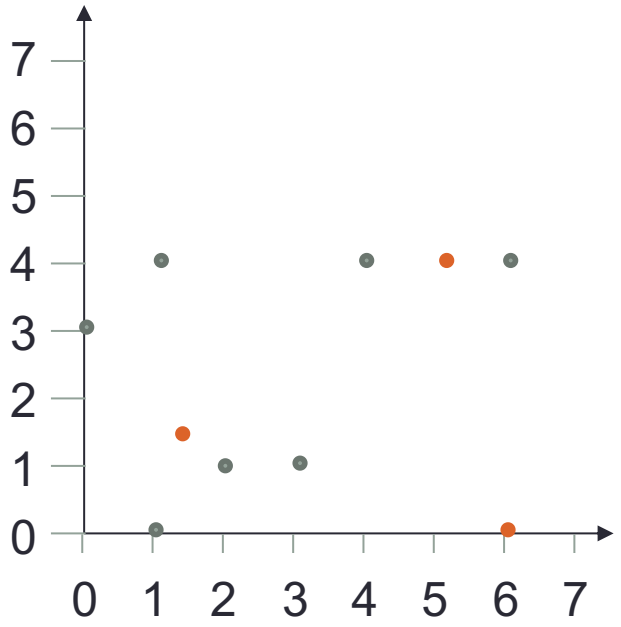
- 9 Beobachtungen (A bis I) mit zwei numerischen Attributen (x, y)

	A	B	C	D	E	F	G	H	I
x	0	1	1	2	3	4	5	6	6
y	3	4	0	1	1	4	4	4	0

- Anwendung von K-Means mit „ungünstigen“ Initialen Clusterzentren







Beispiel – Numerische Lösung I

- 9 Beobachtungen (A bis I) mit zwei numerischen Attributen (x, y)

	A	B	C	D	E	F	G	H	I
x	0	1	1	2	3	4	5	6	6
y	3	4	0	1	1	4	4	4	0

- Ermittlung Distanzmatrix

		A	B	C	D	E	F	G	H	I
C_1	*(1,4)*	1.41	0.00	4.00	3.16	3.61	3.00	4.00	5.00	6.40
C_2	*(1,0)*	3.16	4.00	0.00	1.41	2.24	5.00	5.66	6.40	5.00
C_3	*(4,4)*	4.12	3.00	5.00	3.61	3.16	0.00	1.00	2.00	4.47
Zuordnung		C_1	C_1	C_2	C_2	C_2	C_3	C_3	C_3	C_3

- Ermittlung neuer Clusterzentren
  - C_1 = (0.50, 3.50)
  - C_2 = (2.00, 0.67)
  - C_3 = (5.25, 3.00)

Beispiel – Numerische Lösung II

- 9 Beobachtungen (A bis I) mit zwei numerischen Attributen (x, y) [wdh.]

	A	B	C	D	E	F	G	H	I
x	0	1	1	2	3	4	5	6	6
y	3	4	0	1	1	4	4	4	0

- Ermittlung Distanzmatrix (2. Schritt)

		A	B	C	D	E	F	G	H	I
C_1	*(0.50,3.50)*	0.71	0.71	3.54	2.92	3.54	3.54	4.53	5.52	6.52
C_2	*(2.00,0.67)*	3.07	3.48	1.20	0.33	1.05	3.88	4.48	5.20	4.06
C_3	*(5.25,3.00)*	5.25	4.37	5.20	3.82	3.01	1.60	1.03	1.25	3.09
Zuordnung		C_1	C_1	C_2	C_2	C_2	C_3	C_3	C_3	C_3

- Ermittlung neuer Clusterzentren nicht mehr nötig!
  ⇒ Ergebnis erreicht

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems

- Clusteringverfahren
  - Prinzip und Distanzmaße
  - Übersicht Verfahren
  - K-Means
  - Hierarchische Verfahren
  - Anwendungsbeispiel

- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Vorgehen I

- 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

Vorgehen II

- 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

Beispiel – Numerische Lösung I

- 9 Beobachtungen (A bis I) mit zwei numerischen Attributen (x, y)

	A	B	C	D	E	F	G	H	I
x	0	1	1	2	3	4	5	6	6
y	3	4	0	1	1	4	4	4	0

- Ermittlung Distanzmatrix (2. Schritt)

		B	C	D	E	F	G	H	I
A	*(0,3)*	1.41	3.16	2.82	3.61	4.12	5.10	6.08	6.70
B	*(1,4)*	-	4.00	3.16	3.60	3.00	4.00	5.00	6.40
C	*(1,0)*	-	-	1.41	2.24	5.00	5.66	6.40	5.00
D	*(2,1)*	-	-	-	1.00	3.61	4.24	5.00	4.12
E	*(3,1)*	-	-	-	-	3.16	3.61	4.24	3.16
F	*(4,4)*	-	-	-	-	-	1.00	2.00	4.47
G	*(5,4)*	-	-	-	-	-	-	1.00	4.12
H	*(6,4)*	-	-	-	-	-	-	-	4.00

Beispiel – Numerische Lösung II

- Zusammenfassung der Cluster F, G mit neuem Zentrum (4.50, 4.00)...
  ... und Ermittlung neuer Distanzen

		B	C	D	E	F,G	H	I
A	*(0,3)*	1.41	3.16	2.82	3.61	4.61	6.08	6.70
B	*(1,4)*	-	4.00	3.16	3.60	3.50	5.00	6.40
C	*(1,0)*	-	-	1.41	2.24	5.31	6.40	5.00
D	*(2,1)*	-	-	-	1.00	3.91	5.00	4.12
E	*(3,1)*	-	-	-	-	3.35	4.24	3.16
F,G	*(4.50,4.00)*	-	-	-	-	-	1.50	4.27
H	*(6,4)*	-	-	-	-	-	-	4.00

Beispiel – Numerische Lösung III

- Zusammenfassung der Cluster D, E mit neuem Zentrum (2.50, 1.00)

		B	C	D,E	F,G	H	I
A	*(0,3)*	1.41	3.16	3.20	4.61	6.08	6.70
B	*(1,4)*	-	4.00	3.35	3.50	5.00	6.40
C	*(1,0)*	-	-	1.80	5.31	6.40	5.00
D,E	*(2.50,1.00)*	-	-	-	3.61	4.61	3.64
F,G	*(4.50,4.00)*	-	-	-	-	1.50	4.27
H	*(6,4)*	-	-	-	-	-	4.00

- Zusammenfassung der Cluster A, B mit neuem Zentrum (0.50, 3.50)

		C	D,E	F,G	H	I
A,B	*(0.50,3.50)*	3.53	3.20	4.03	5.52	6.51
C	*(1,0)*	-	1.80	5.31	6.40	5.00
D,E	*(2.50,1.00)*	-	-	3.61	4.61	3.64
F,G	*(4.50,4.00)*	-	-	-	1.50	4.27
H	*(6,4)*	-	-	-	-	4.00

Beispiel – Numerische Lösung IV

- Zusammenfassung der Cluster F, G, H mit neuem Zentrum (5.00, 4.00)

		C	D,E	F,G,H	I
A,B	*(0.50,3.50)*	3.53	3.20	4.53	6.51
C	*(1,0)*	-	1.80	5.66	5.00
D,E	*(2.50,1.00)*	-	-	3.91	3.64
F,G,H	*(5.00,4.00)*	-	-	-	4.12

- Zusammenfassung der Cluster C, D, E mit neuem Zentrum (2.00, 1.00)

		C,D,E	F,G,H	I
A,B	*(0.50,3.50)*	2.92	4.53	6.51
C,D,E	*(2.00,1.00)*	-	4.24	4.12
F,G,H	*(5.00,4.00)*	-	-	4.12

- Zusammenfassung der Cluster A,B,C,D,E mit neuem Zentrum (1.40, 1.80)

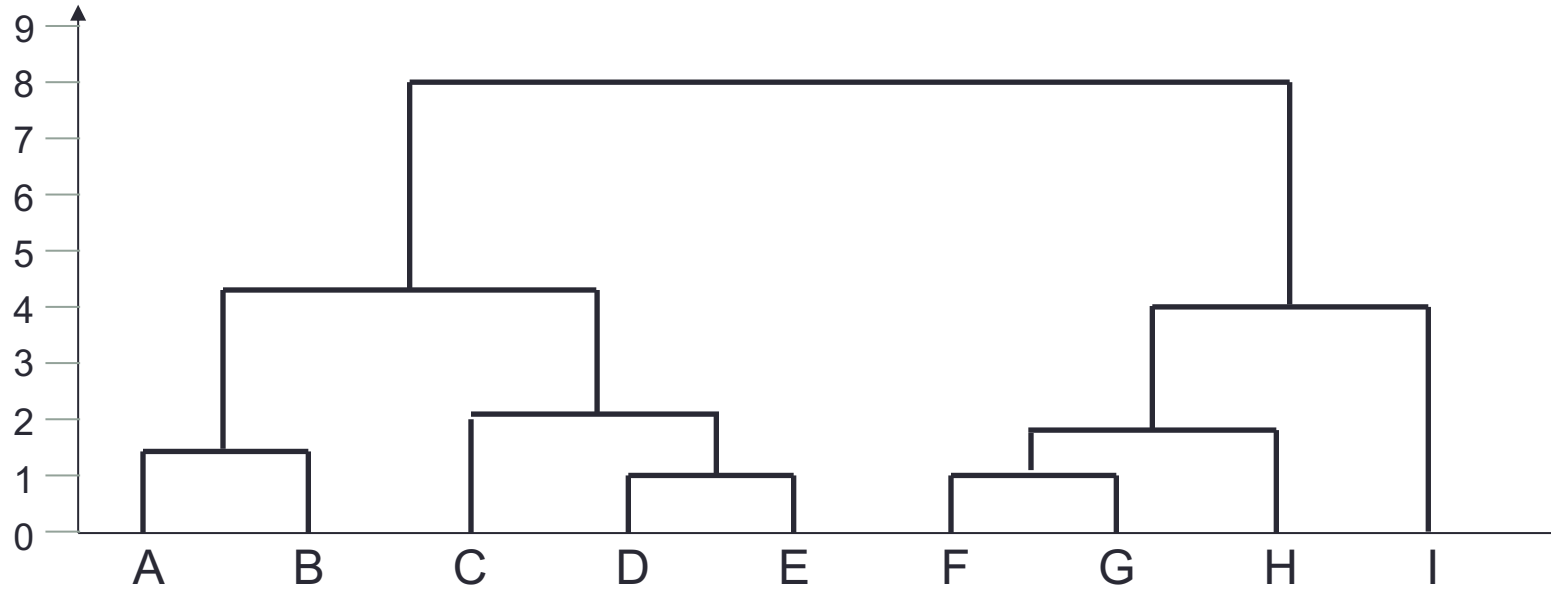
		F,G,H	I
A,B,C,D,E	*(1.40,1.80)*	4.21	4.94
F,G,H	*(5.00,4.00)*	-	4.12

Beispiel – Numerische Lösung IV

- Zusammenfassung der Cluster F,G,H,I mit neuem Zentrum (5.25, 2.40)

		F,G,H,I
A,B,C,D,E	*(1.40,1.80)*	3.90

- Zusammenfassung der Cluster A,B,C,D,E,F,G,H,I
- Dendrogramm



Dendrogramm

- Idee
  - Visualisiert sämtlich Zusammenfassungen von Clustern...
    ... beim Ermitteln des hierarchisch agglomerativen Clusterings
  - Höhe der Verbindungen repräsentiert Abstand der Cluster bei Kombination

- Eigenschaften des Dendrogramms
  - Dendrogramm ist Binärbaum
  - Blätter repräsentieren einzelne Beobachtungen
  - Nachfolger eines Knoten sind Cluster in die er aufgespalten wird
  - Kantenlänge repräsentiert Distanz der Cluster

- Vorteil des agglomerativen Clusterings (Dendrogramm) gegenüber K-Means
  - Optimale Clusterzahl aus Dendrogramm “ablesbar“:
    Clustering, so dass „hohe“ Distanzen zu getrennten Clustern führen

Klassifikation mit Clustering Konzepten

- Algorithmus: K-Nearest-Neighbor Klassifikation
- Gegeben:
  - Trainingsdatensatz T
- Vorgehen zur Klassifikation einer Beobachtung g_i:
  - Menge S: k ($k\ modulo\ 2 = 1$) 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 das Mittelwerts der k-Nearest-Neighbors

Anmerkungen Clustering

- Hoher (intellektueller) Aufwand vor Anwendung des Algorithmus
  - Auswahl geeigneter Attribute
  - Transformation von Wertebereichen
  - Kombination von Attributen mit unterschiedlichen Skalenniveaus

- Hoher (intellektueller) Aufwand nach Anwendung des Algorithmus
  - Kleine Anpassungen führen zu deutlich unterschiedlichen Ergebnissen
  - Zur Prüfung Stabilität mehrere Anwendungen des Algorithmus nötig
  - Güte nicht über Kennzahl ableitbar,...
    ... bzw. Kennzahl muss pro Anwendungsfall definiert werden

- Interessant
  - Probabilistische Ansätze (d.h. unscharfe Zuordnung zu Clustern)
  - Potential zum besseren Verständnis der Kunden, ...

Agenda

- Einführung
- Deskriptive Methoden zur Datenexploration
- Datenqualität
- Klassifikation
- Recommender Systems

- Clusteringverfahren
  - Prinzip und Distanzmaße
  - Übersicht Verfahren
  - K-Means
  - Hierarchische Verfahren
  - Anwendungsbeispiel

- Stream Mining
- Social Network Analysis
- Technische Lösungen
- Datenschutz und gesellschaftliche Aspekte

Herausforderungen im Handel

- Aufgabe (über die Zeit konstant)
  - Identifikation von Sortiment und Kunden
  - Zusammenbringen Kunde und Sortiment
  - Sekundär: Maximierung des Erlöses
- Trotzdem: „Händlersterben“
  - Versandhändler
    Quelle, Neckermann, übrige konsolidiert
  - Warenhausketten
    Karstadt, Galeria Kaufhof
- Aber: Viele neue Wettbewerber
  Neue Händler fokussiert auf einzelne Marken, Sortimente oder Kundengruppen
- Ökonomisch nicht plausibel
  - Synergieeffekte
  - Cross Selling Potentiale

- Hier: Kann verbesserte Abstimmung zwischen Sortiment und Kunde helfen?

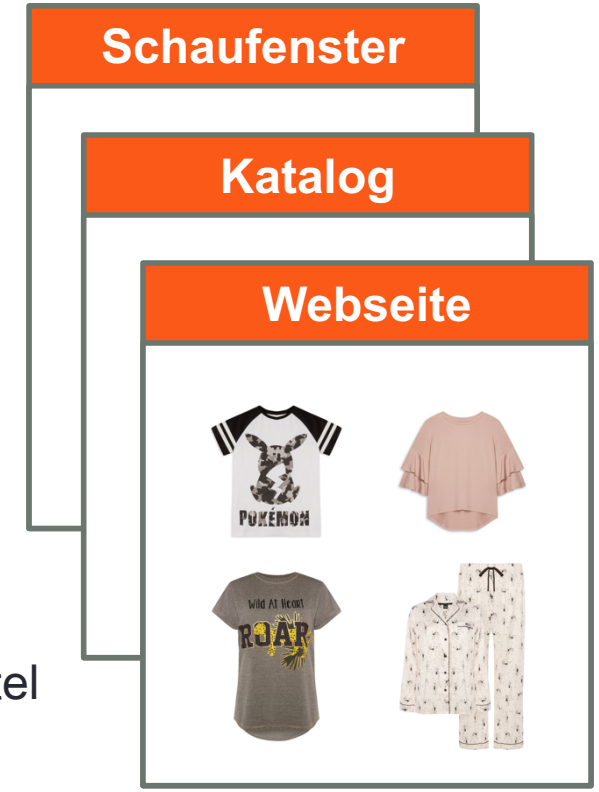
Vom Sortiment zum Kunden – Kundenscoring

- Empirische Betrachtungsweise

- Vorgehen
  - Messung der „besten“ Kunden über historische Performancedaten
  - Werbemittel geht an „beste“ Kunden

- Aufbau von Scoringmodellen
  - Identifikation von Kennzahlen pro Kunde (Retourenquote, Bruttoumsatz, ...)
  - Ableitung eines Regressionsmodells zur Vorhersage der Kaufwahrscheinlichkeit
  - Adressierung der Kunden mit höchster Kaufwahrscheinlichkeit (Score) durch Werbemittel

- Nutzen
  Reduktion der Streuverluste



Funktion eines Regressionsmodells

bestellte

Boxershorts
und
Socken

Retourenquote (rq): 0 %

Bruttonachfrage (bn): 50 €

Response Katalog (rk): 100 %

- Im Regressionsmodell (auf Basis historischer Daten)
  Score = $-0.5 \cdot rq + 1.0 \cdot bn + 0.3 \cdot rk$
  ⇒ Kunde würde sicher Winterkatalog erhalten!

- Kernfokus Regressionsmodell
  - Identifikation „signifikanter“ Einflussfaktoren
  - Richtung des Zusammenhangs

Vom Sortiment zum Kunden – Treatmentvergleiche

- Empirische Betrachtung im Labor

- Vorgehen
  - Hypothese bzgl. Unterschieden in Werbemitteln
  - Test der Hypothese über Treatmenteffekte
  - Anpassung künftiger Werbemittel gemäß Hypothese

- Ablauf Hypothesentest
  - Identifikation potentieller Einflussfaktoren
  - Erstellung von Werbemitteln mit Unterschieden in den Einflussfaktoren
  - Einladen guter Kunden (homogen!)
  - Messen der Wirkung Einflussfaktor auf Kennzahlen

- Nutzen
  - Optimierung des Werbemittels möglich

Funktion eines Hypothesentests

Treatment 1 Treatment 2

Artikel auf rotem Hintergrund verkaufen sich besser

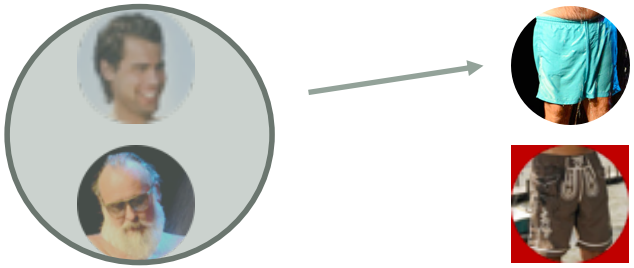
- Probanden machen „Erfahrung“ mit geringen Unterschieden
- Reaktion auf “Erfahrung“ gemessen (z.B. über Nachfrage)
- Beeinflusst Erfahrungsänderung Verhalten, wird Änderung im Feld umgesetzt

Zusammenfassung traditionelles Vorgehen

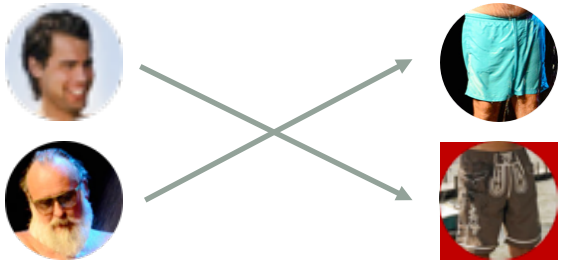
- Regressionsmodell
  - Findet monotonen Zusammenhänge zwischen verschiedenen Größen
  - Lässt sich von empirischen Beobachtungen ableiten
  - Reduziert Streuverluste
- Hypothesentest
  - Findet qualitative Unterschiede zwischen Treatments
  - Lässt sich im Labor leicht/kostengünstig durchführen
  - Optimiert Werbemittel
- Anwendung in der Praxis
  - Kundenscoring (über Regressionsmodell)
  - Treatmentvergleiche (über Hypothesentests)
- Herausforderung
  Würden Ansätze funktionieren hätten traditionelle Händler kein Problem...
- Offene Fragen
  Kann das Sortiment passender zugeschnitten werden?

Verbesserung der Kundenansprache

- Bisher
  - Rückkopplung nur „der Kunde“ kauft Sortiment XY
  - Nachfrage nach „dem Artikel“ steigerbar durch Maßnahme XY

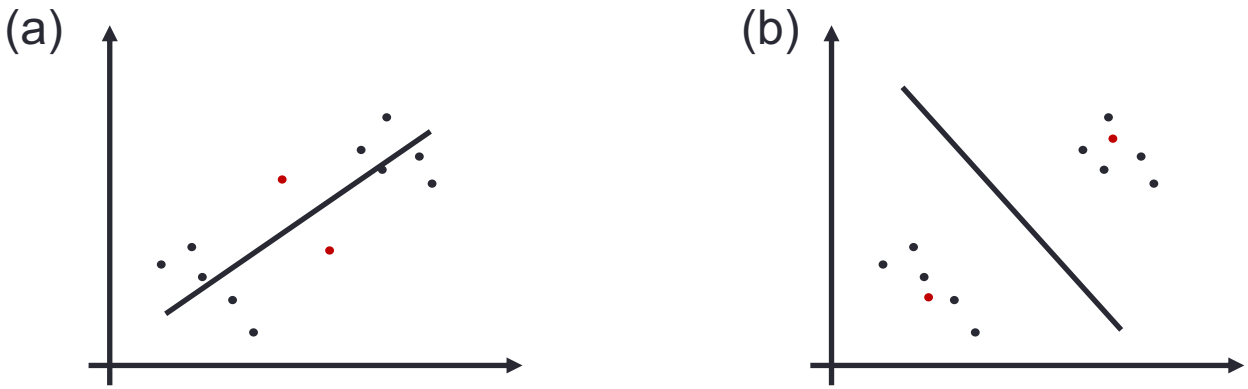


- Jetzt
  - Unterteilung der Kunden nach Präferenzen (Clustering)
  - Zuordnung des passenden Sortiments zu Kunden



Clustering – Bewertung

- Zwei Clusterings auf identischen Daten

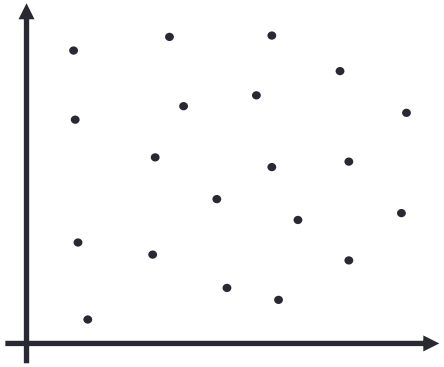


- Welches ist „besser“? – Natürlich (b)

- Formal (Silhoutenkoeffizient)
  - Ermitteln der durchschnittlichen „Silhoute“...
    ... d.h. dem mittleren Abstand der Datenpunkte vom Clusterzentrum
  - „Silhoute“ bei (b) deutlich geringer als bei (a)

Clustering – Bewertung in der Praxis

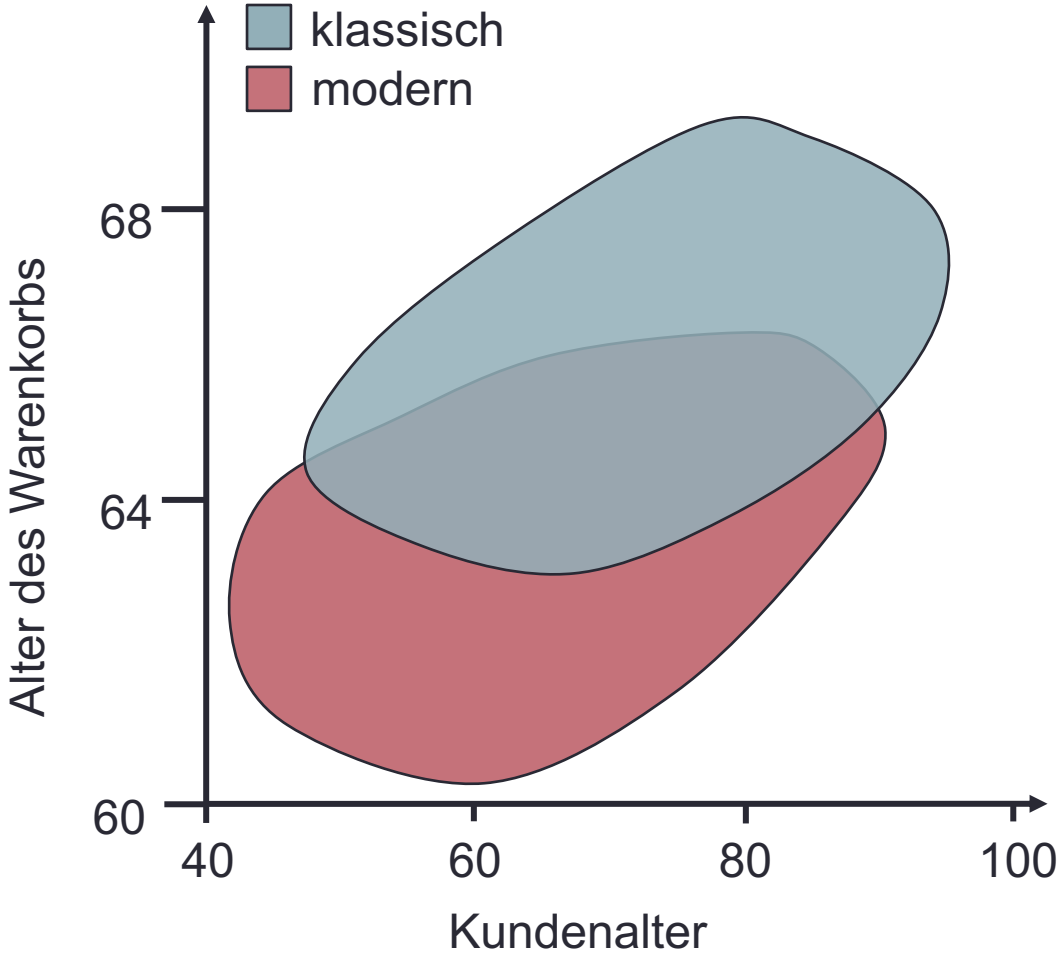
- In der Praxis sind Cluster meist nicht offensichtlich



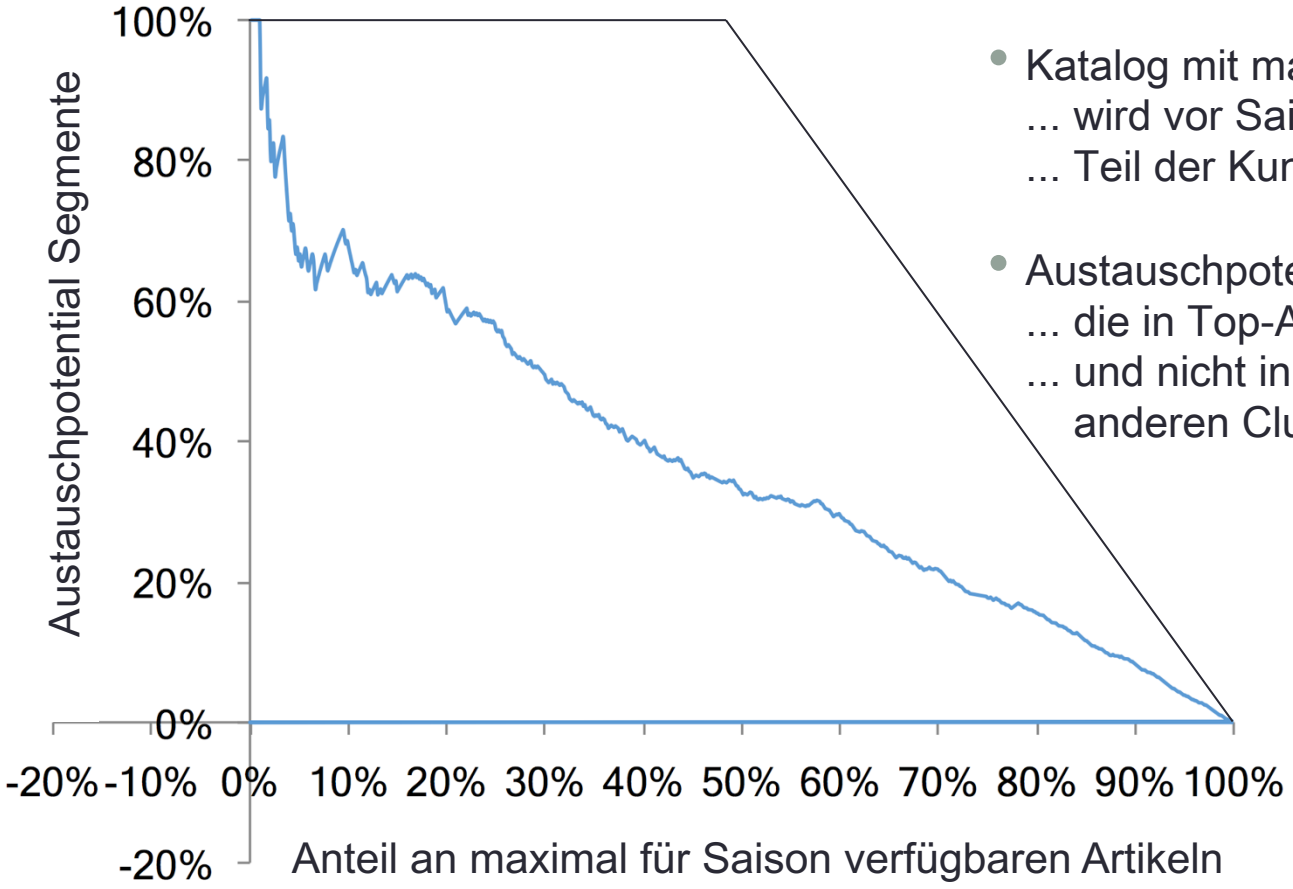
- Offene Frage: Wie lässt sich bestes Clustering bestimmen?
- Einfache Antwort: Nicht mit dem Silhouttenkoeffizient!
- Ausführliche Antwort (siehe nächste Folien):
  - Mit ökonomischer Sichtweise: Einfluss auf Nachfrage...
  - ... auf Basis der Segmente

Clustering – Primär trennende Attribute

- Alter als Maß der Modizität



Austauschpotential vs. Sortimentgröße



- Ermittlung über „Vortest“:
- Katalog mit max. Sortiment...
  ... wird vor Saison...
  ... Teil der Kunden angeboten
- Austauschpotential sind Artikel...
  ... die in Top-Artikeln vom einen Cluster
  ... und nicht in Top-Artikeln des anderen Clusters liegen

Uplift bei Ausnutzung des Austauschpotentials

Test der Kundensegmentierung

Modern

Auflage:	60.000
∅ Preis:	34,30 €
Laura Kent:	29,8 %
Paola Classic:	12,5 %

Klassisch

Auflage:	60.000
∅ Preis:	31,76 €
Laura Kent:	10,1 %
Paola Classic:	25,9 %

Basis

Auflage:	120.000
∅ Preis:	32,90 €
Laura Kent:	19,6 %
Paola Classic:	18,3 %

- Organisatorisches
  - Aussandtermin: 23.12.2016
  - Seiten: 126
- Angebotene Artikel
  - 20 % Austauschpotential (nach Ranking)
  - 80 % Sonstige (Überschneider)
- Übrige Eigenschaften
  - Identische Verkaufsförderung
  - Layout ähnlich
- Wahl der Adressaten
  - Hälfte Modisch / Klassisch... ... bzw. Basis
  - Hälfte pro Segment erhält... ... passenden Katalog bzw. Basis

Austauschpotential in den Katalogen

- „Moderne Artikel“: Von modernen Kunden häufig, sonst kaum gekauft

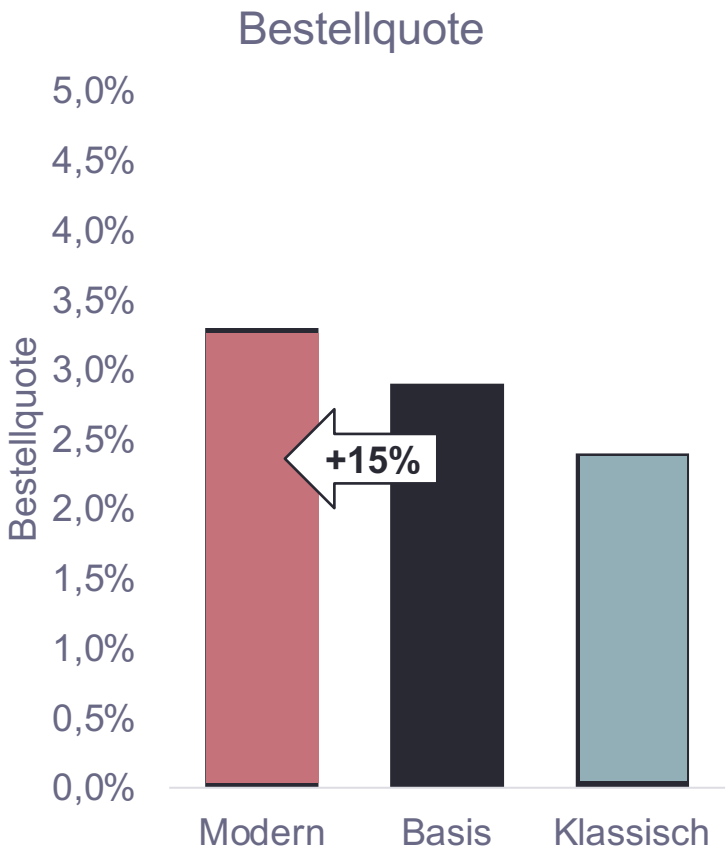


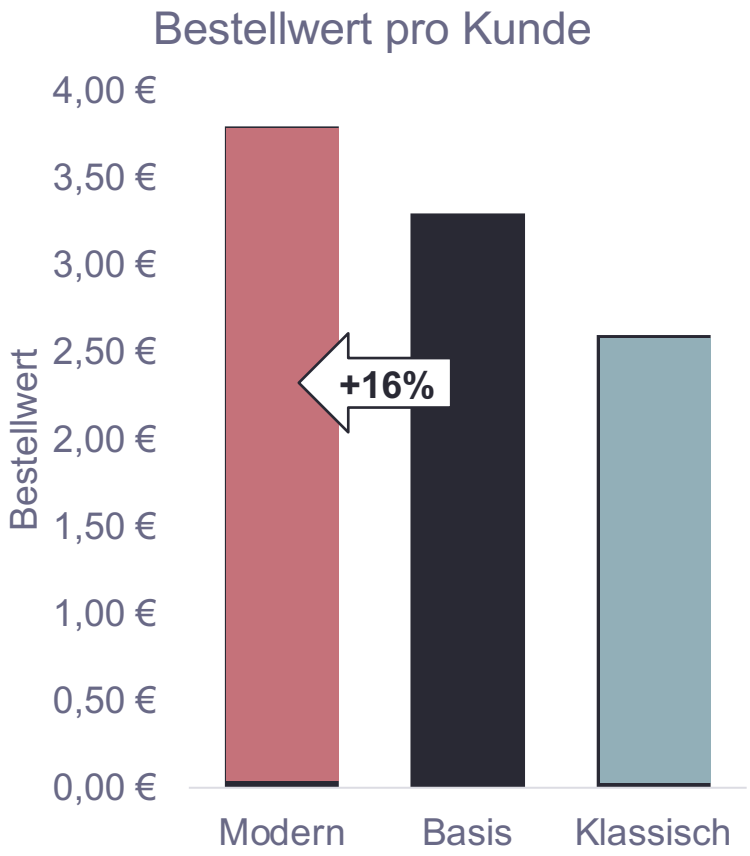

- „Klassische Artikel“: Von klassischen Kunden häufig, sonst kaum gekauft



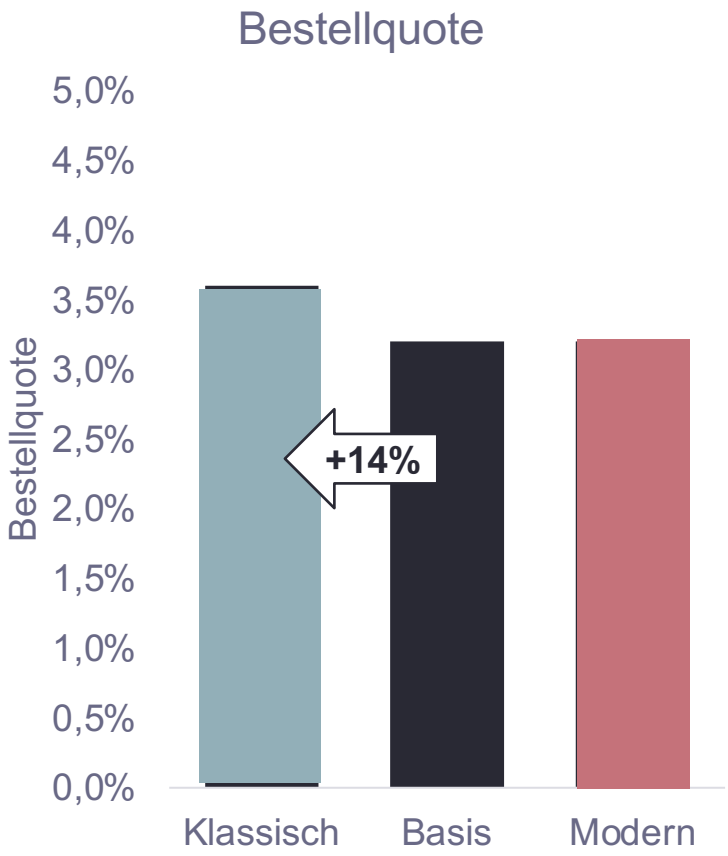


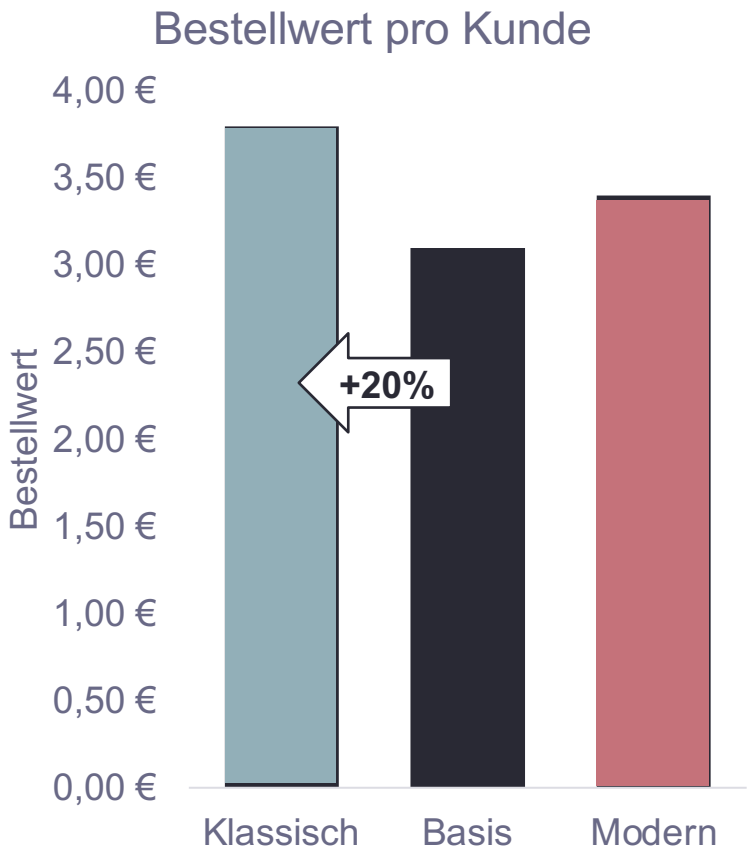
Einfluss auf Nachfrage – Modische Kunden



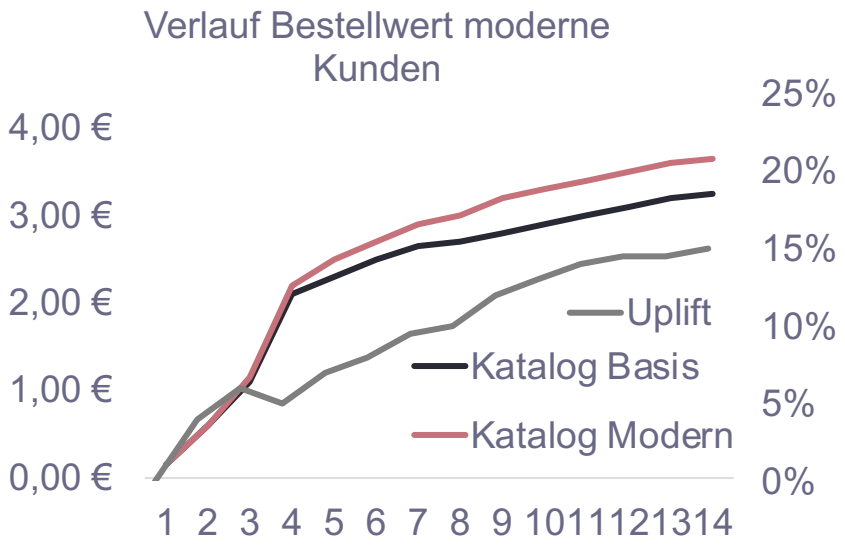


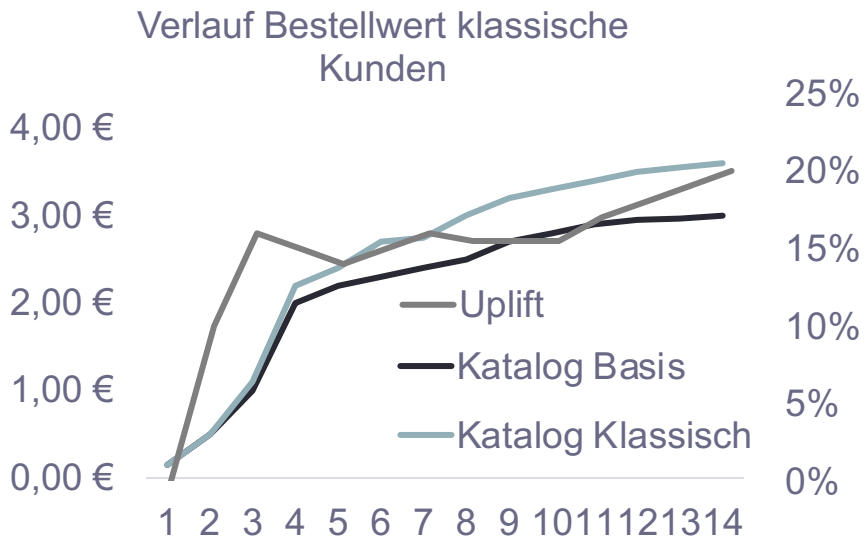
Einfluss auf Nachfrage – Klassische Kunden





Zeitliche Entwicklung





- Uplift (Quotient aus Bestellwert zwischen Basis und Segmentspezifischem Katalog) steigt über die Zeit
- Höhere Attraktivität des Sortiments begünstigt längere Laufzeit

Vergleich Überschneider zu Austauschartikeln

Stücknachfrage	**Segment Modern**				**Segment Klassisch**		
Katalog	**Modern**	**Basis**	**Abw.**		**Klassisch**	**Basis**	**Abw.**
Überschneider	22,05	21,71	2 %		23,14	21,23	9 %
Austausch	17,61	10,59	66 %		16,68	9,84	69 %
Gesamt	21,32	19,39	10 %		21,71	18,96	14 %

- Austauschartikel deutlich besser als insgesamt schlechteste Gute
- Austauschartikel erhöhen auch mittleren Erfolg der Überschneider