

DATA SCIENCE 2

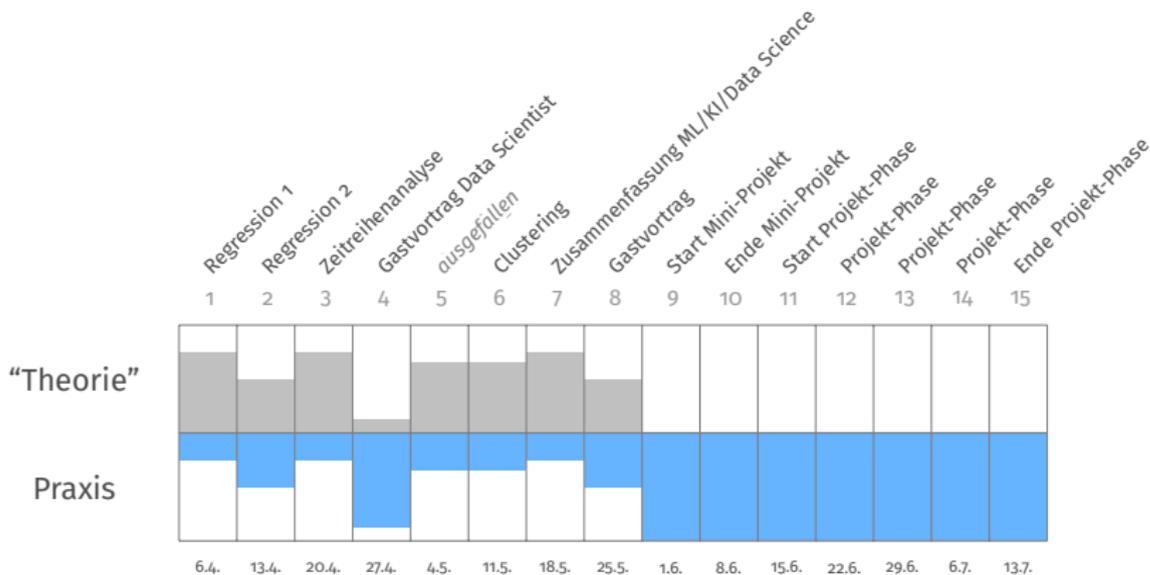
VORLESUNG 5 - INTRO

PROF. DR. CHRISTIAN BOCKERMANN

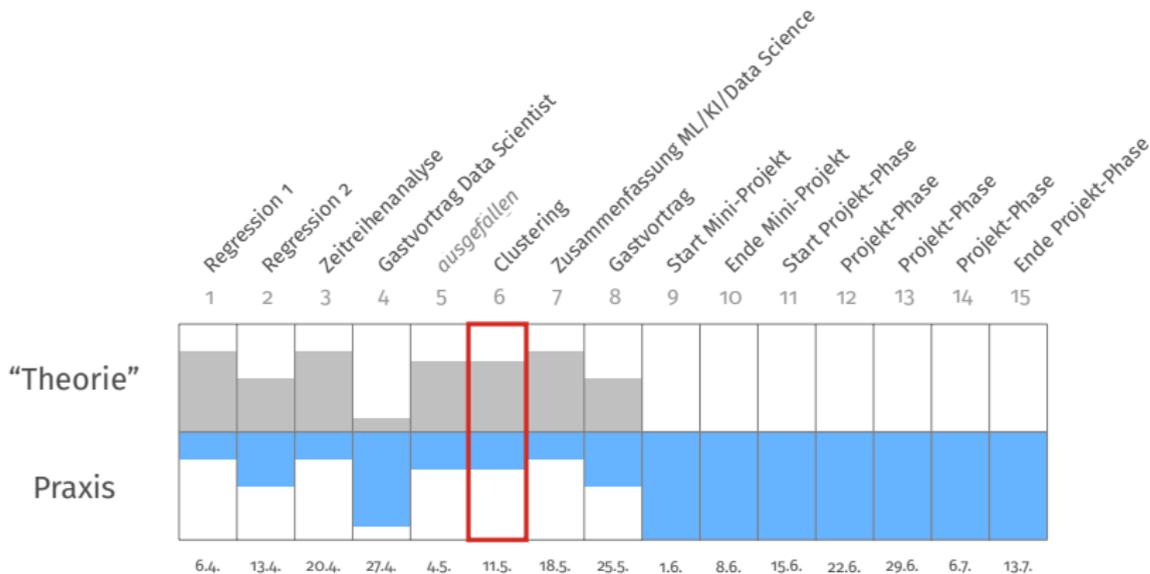
HOCHSCHULE BOCHUM

SOMMERSEMESTER 2021

Themen der Vorlesung



Themen der Vorlesung



Wiederholung - Clustering

Clustering sucht Aufteilung von Daten in ähnliche Gruppen

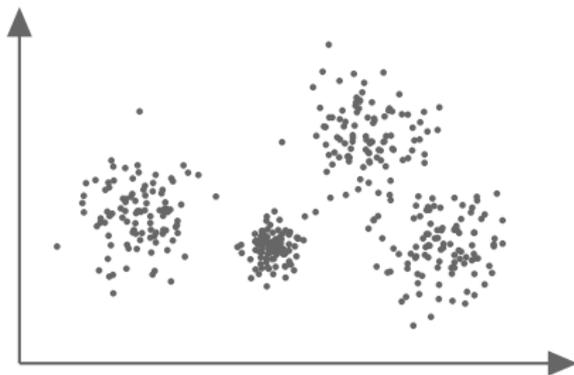
- Datenmenge \mathbf{X} von Beispielen (keine Klassen gegeben!)
- Parameter k zu findender Gruppen
- Abstandsmaß $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$
- Qualitätsfunktion q

Ziel:

- Abstand *innerhalb* der Gruppen soll minimiert, Abstand *zwischen* den Gruppen soll maximiert werden

Beispiel: Clustering

Sei $\mathbf{C} = C_1, \dots, C_k$ eine Aufteilung der Daten X (ein *Clustering*)

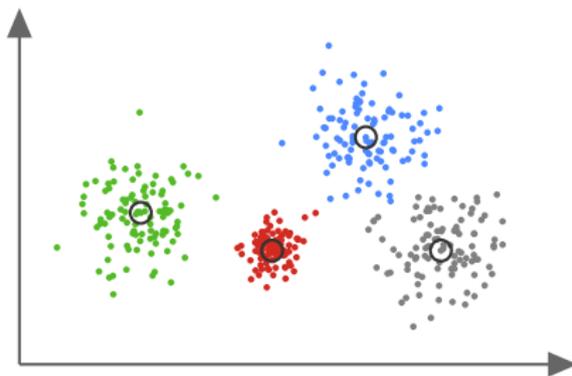


Qualitätsfunktion: (Innere Abstände)

$$q_{inner}(\mathbf{C}) = \sum_{i=1}^k \sum_{x \in C_i} d(x, \bar{c}_i) \quad , \text{ mit } \bar{c}_i \text{ Zentrum von } C_i$$

Beispiel: Clustering

Sei $\mathbf{C} = C_1, \dots, C_k$ eine Aufteilung der Daten X (ein *Clustering*)



Clustering auf Datenpunkten mit $k = 4$. Die schwarzen Kreise markieren jeweils das Zentrum \bar{c}_i des jeweiligen Cluster C_i .

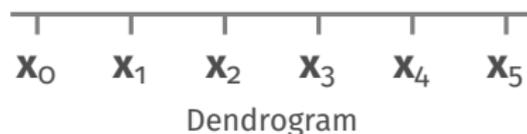
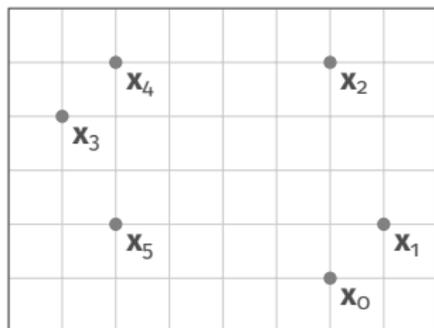
Qualitätsfunktion: (Innere Abstände)

$$q_{inner}(\mathbf{C}) = \sum_{i=1}^k \sum_{x \in C_i} d(x, \bar{c}_i) \quad , \text{ mit } \bar{c}_i \text{ Zentrum von } C_i$$

Einordnung von Clustering-Verfahren

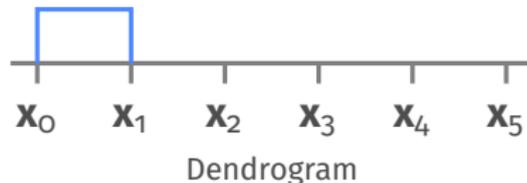
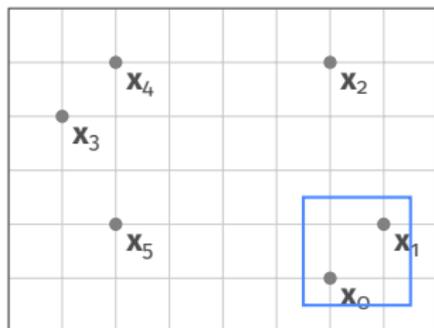
- Hierarchisches Clustering, agglomerativ/divisiv
- Iterative Verfahren (z.B. [k-Means](#))
- Dichte-basiertes Clustering (z.B. DBScan)
- Stochastische Verfahren
- Meta-Daten Basierte Verfahren

Hierarchisches Clustering (agglomerativ)



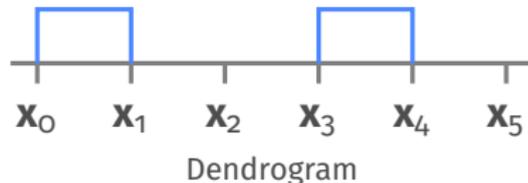
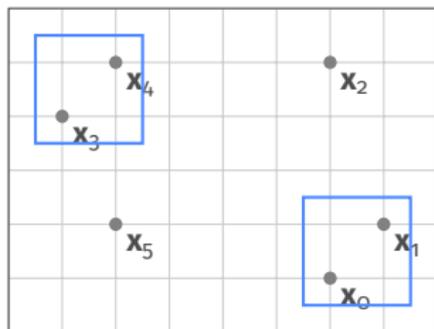
- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die **nächstgelegenen Cluster** zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)



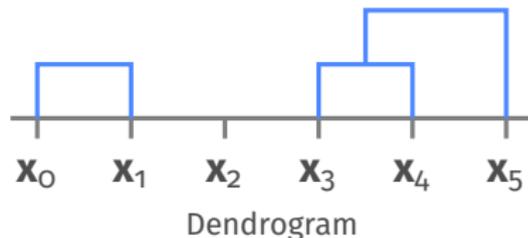
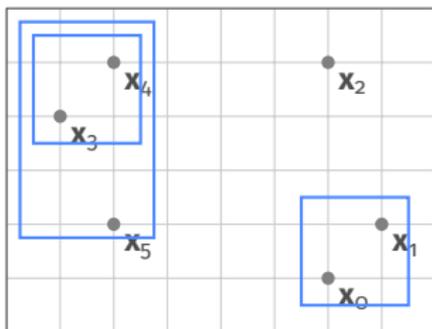
- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die **nächstgelegenen Cluster** zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)



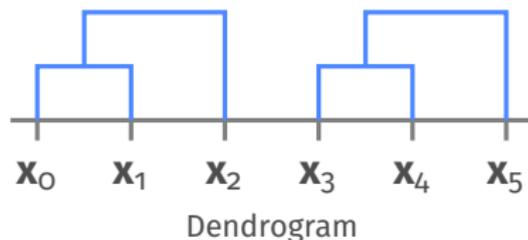
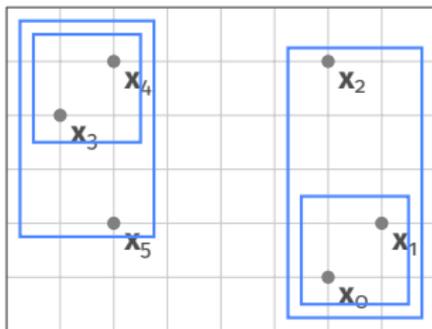
- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die **nächstgelegenen Cluster** zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)



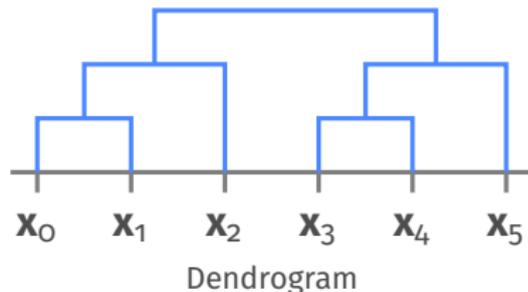
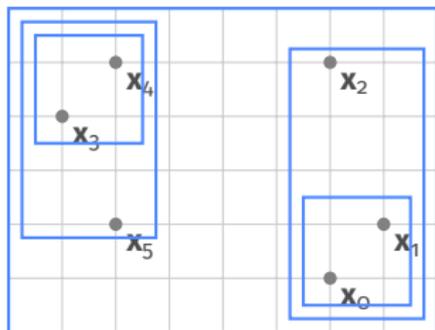
- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die **nächstgelegenen Cluster** zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)



- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die **nächstgelegenen Cluster** zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)



- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die **nächstgelegenen Cluster** zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

k-Means Algorithmus

- Distanz-basiertes, iteratives Clustering-Verfahren
- Erzeugt k disjunkte Teilmengen von \mathbf{X}
- k ist Benutzer-Parameter
- Distanzfunktion wird auch von Benutzer gewählt

Algorithmus: k-Means

1. Wähle k zufällige Clusterpunkte c_1, \dots, c_k aus \mathbf{X}
2. Ordne jedes $\mathbf{x} \in \mathbf{X}$ dem nächstgelegenen c_j zu, d.h.

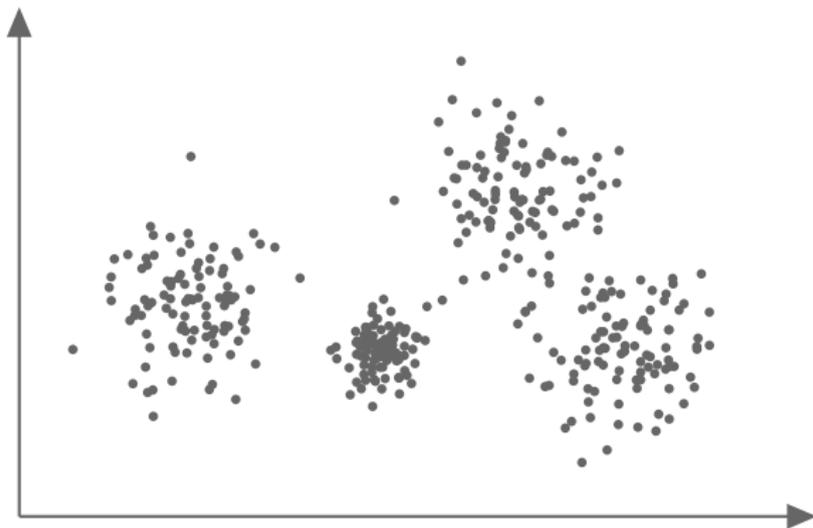
$$\mathbf{C}_j = \{ \mathbf{x} \in \mathbf{X} \mid \mathbf{x} \text{ am nächsten an } \mathbf{c}_j \}$$

3. Berechne neue Clusterpunkte

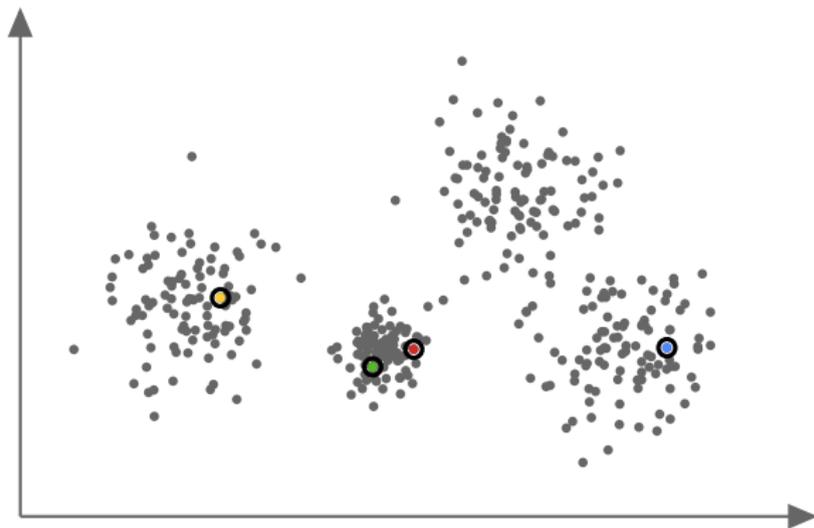
$$\bar{\mathbf{c}}_j = \frac{1}{|\mathbf{C}_j|} \sum_{\mathbf{x}_j \in \mathbf{C}_j} \mathbf{x}_j$$

Wenn $\bar{\mathbf{c}}_1, \dots, \bar{\mathbf{c}}_k \simeq \mathbf{c}_1, \dots, \mathbf{c}_k$, dann STOP, sonst springe zu **2.**
mit den neuen Punkten $\bar{\mathbf{c}}_1, \dots, \bar{\mathbf{c}}_k$

Beispiel: k-Means

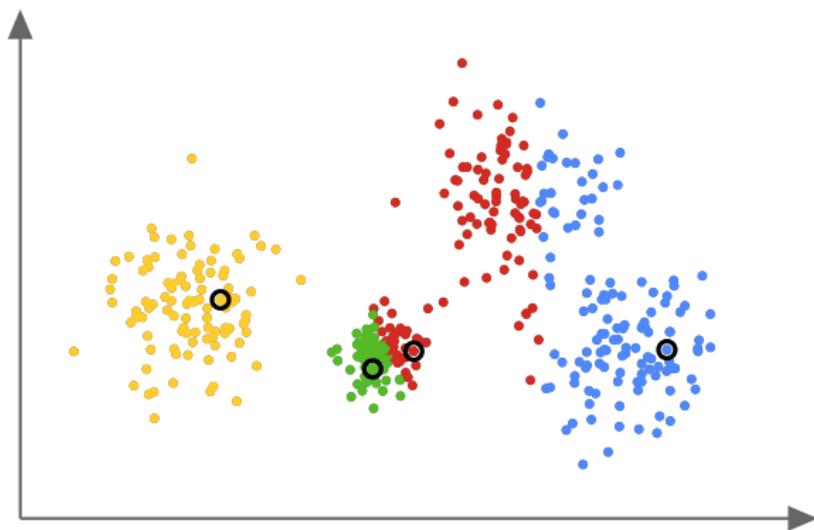


Beispiel: k-Means



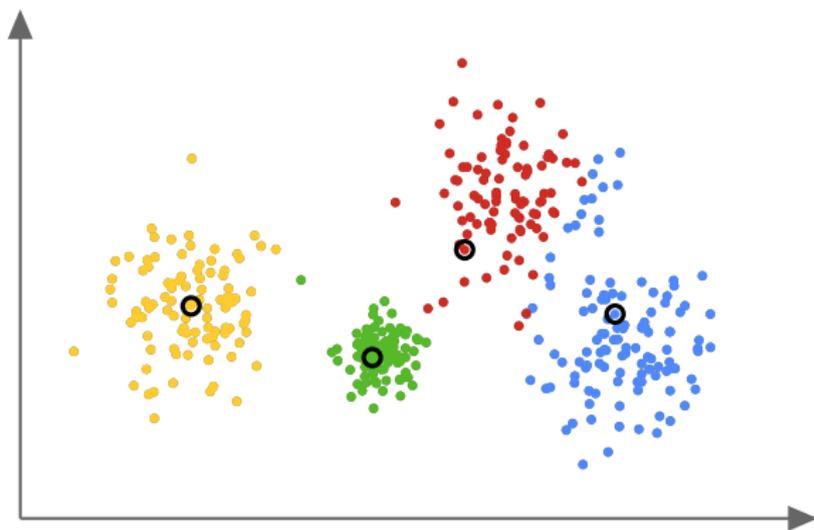
1. Wähle k zufällige Cluster-Mittelpunkte $\mathbf{c}_1, \dots, \mathbf{c}_k$

Beispiel: k-Means



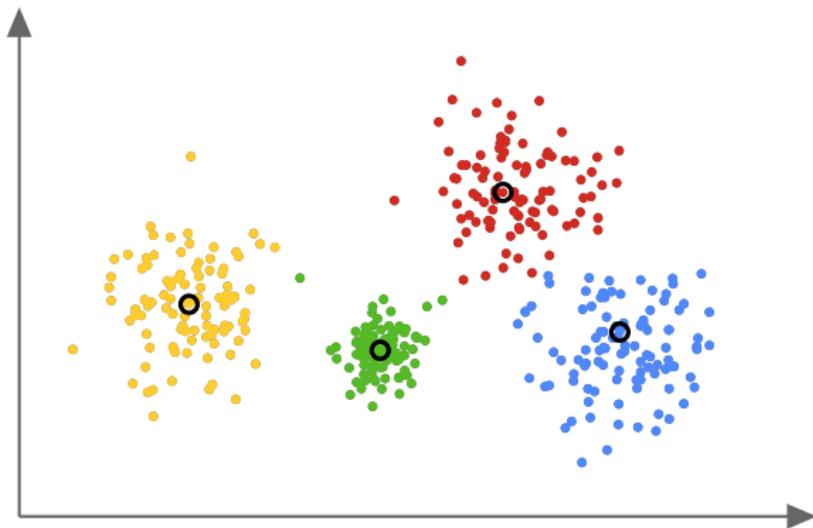
2. Ordne jedem Punkt seinen nächsten Cluster-Mittelpunkt zu

Beispiel: k-Means



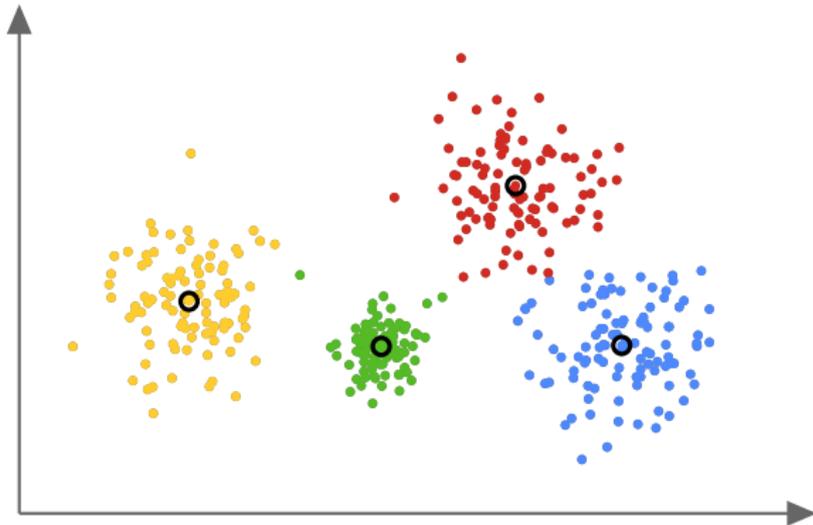
3. Für die Cluster neue Mittelpunkte berechnen, Punkte zuordnen

Beispiel: k-Means



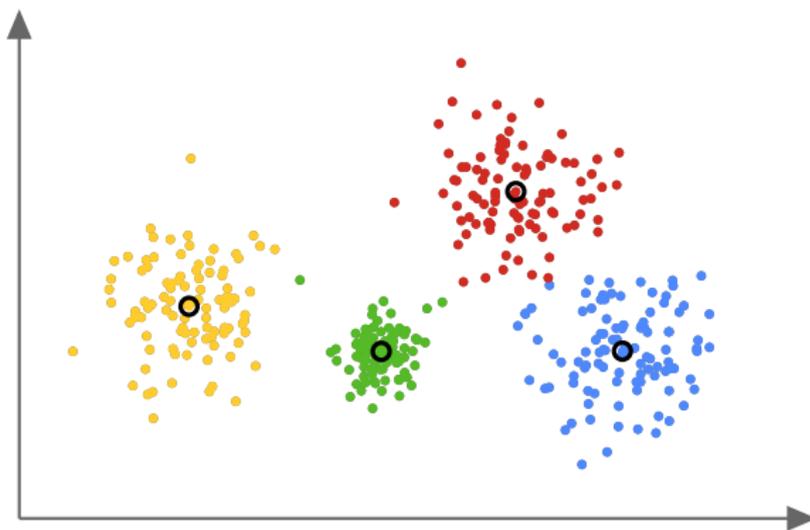
Schritte 2. und 3. wiederholen bis Cluster-Mittelpunkte stabil

Beispiel: k-Means



Schritte 2. und 3. wiederholen bis Cluster-Mittelpunkte stabil

Beispiel: k-Means



Schritte 2. und 3. wiederholen bis Cluster-Mittelpunkte stabil

Clustering - e-Commerce

Clustering von Produkten

- Produktbeschreibungen als Dokument
- Ähnlichkeit von Produkten über Beschreibungen

Ziel: Finde Gruppen ähnlicher Produkte!

Clustering von Produkten

- Produktbeschreibungen als Dokument
- Ähnlichkeit von Produkten über Beschreibungen

Ziel: Finde Gruppen ähnlicher Produkte!

Anwendungen:

- Empfehlen ähnlicher Produkte
- Replacement-Produkte für ausverkaufte Artikel finden!

Kundenprofile – e-Commerce

- Kunden-Daten (Geschlecht, Alter, Wohnort,...)
- Jeder Kunde hat mehrere Produkte gekauft

Kundenprofile – e-Commerce

- Kunden-Daten (Geschlecht, Alter, Wohnort,...)
- Jeder Kunde hat mehrere Produkte gekauft
- Kunde = Dokument mit Worten aus allen gekauften Produkten

Ziel: Finde Gruppen ähnlicher Kunden!

Kundenprofile – e-Commerce

- Kunden-Daten (Geschlecht, Alter, Wohnort,...)
- Jeder Kunde hat mehrere Produkte gekauft
- Kunde = Dokument mit Worten aus allen gekauften Produkten

Ziel: Finde Gruppen ähnlicher Kunden!

Anwendungen:

- Gruppen für gemeinsame Marketing-Maßnahmen
- Bilde Gruppen für Suchprofile (Ranking)

Vorschau: Vorlesung 6

- Zusammenfassung/Überblick Maschinelles Lernen
- Diskussion: [Wie werde ich Data Scientist?](#)