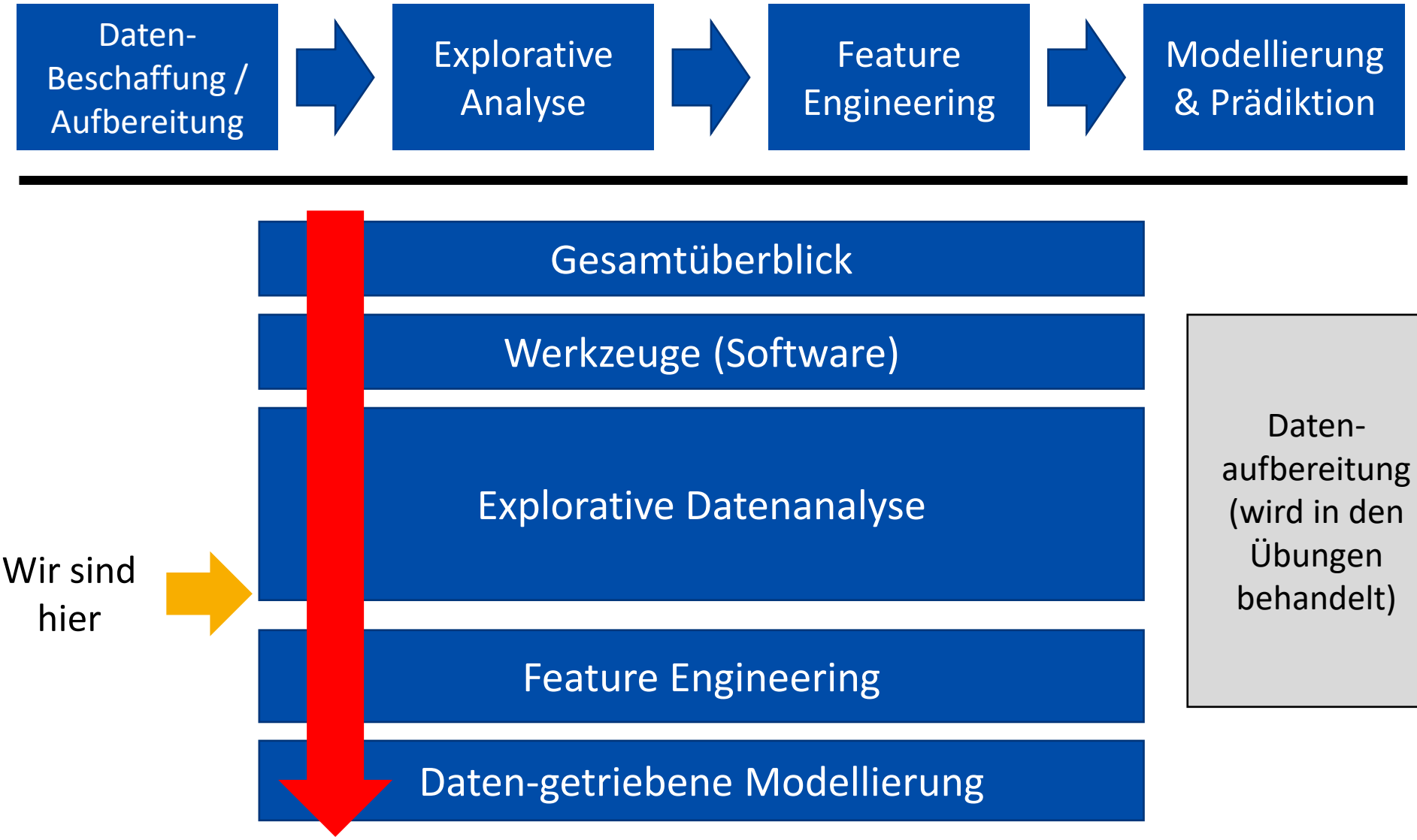
Einführung in Data Science

Unser Plan für heute:

1. Clustervalidierungsverfahren

Klausurtermin: 24. Juli, 9:00-10:15 Uhr

Data Science



Curriculum

1. Grundbegriffe / Überblick
2. Zentrale Softwarebibliotheken

3. Univariate explorative Analyse (EDA) Visualisierung (Teil 1)
4. Visualisierung (Teil 2), Multivariate explorative Analyse (Teil 1)
5. Multivariate explorative Analyse (Teil 2)
6. Dimensionsreduktion (Teil 1): PCA
7. Dimensionsreduktion (Teil 2): MDS, Isomap
8. Clustering: K-Means, HCA
9. Clustervalidierung

10. Probeklausur

11. Feature Engineering, Datengetriebene Modellierung (Teil 1)
12. Datengetriebene Modellierung (Teil 2)

Überblick / Begriffe

Explorative Analyse (EDA)

Feature Engineering & Modellierung

Clustervalidierungsverfahren

Beobachtung

- Cluster-Analyseverfahren finden immer Cluster in den Daten.
- Welches Clustering (Partitionierung) ist sinnvoll, welches nicht?

Clustervalidierungsverfahren

- bewerten Qualität von Clusterings (Partitionen)
  → ermöglichen das Einstellen von Hyperparametern (Beispiel: Anzahl *K* der Cluster bei *K-Means*)
- viele Verfahren publiziert; keine endgültige Empfehlung für eine Methode möglich[1]
- aktueller Forschungsgegenstand

1) Everitt BS, Landau S, Leese M, Stahl D. Cluster Analysis, Wiley 2011

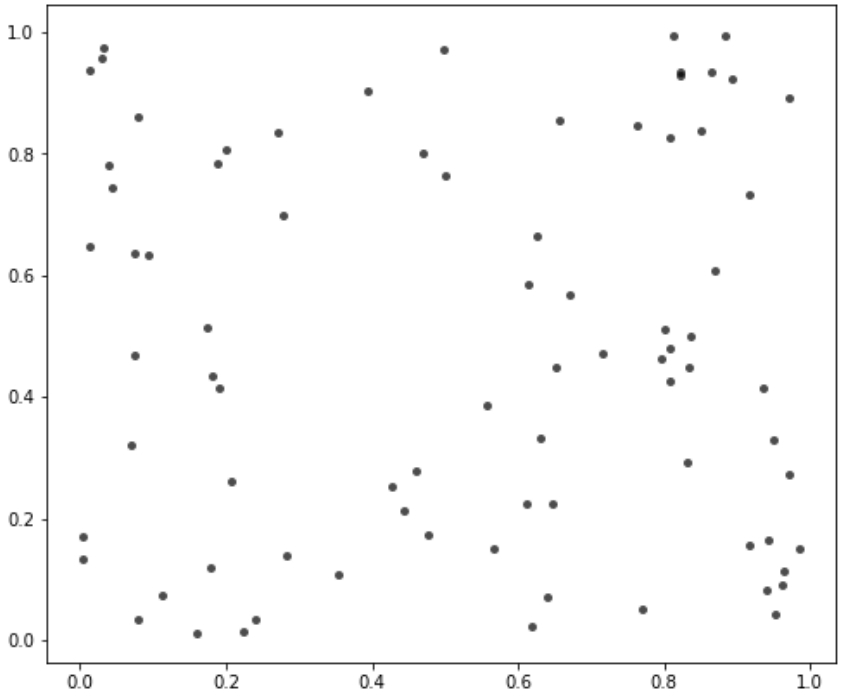
Clustervalidierung

Interaktivität

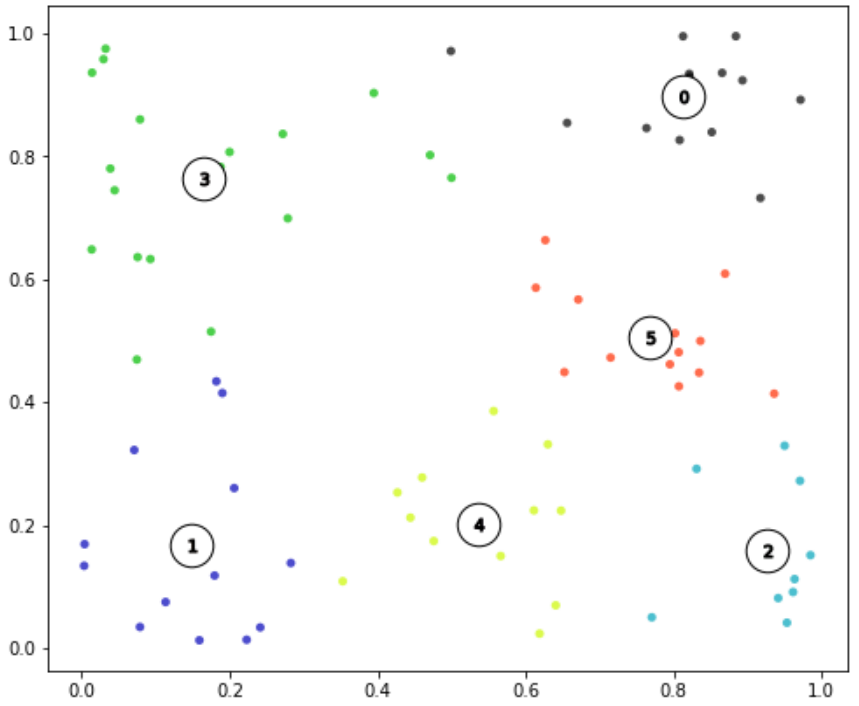
F

1. Können Sie Cluster in den dargestellten Daten finden?
2. Wieviele Cluster haben Sie identifiziert?

Clusteralgorithmen finden immer Cluster, auch wenn die Daten keine Cluster enthalten.



Zufällige Daten (uniform verteilt)



K-Means (6 Clustern)

Clustervalidierung | Arten von Verfahren

1. **Tests auf Abwesenheit von Clusterstruktur**
   - häufig in Forschungskontexten, selten in der Praxis
   - Tests gegen typische Nullhypothesen
     a) Daten liegen zufällig verteilt (uniform) im Merkmalsraum (Uniformitätshypothese)
     b) Daten wurden aus unimodaler Verteilung gezogen (Unimodale Nullhypothese)

→ Behandeln wir hier nicht weiter (da selten in der Praxis genutzt).

2. **Externe Clustervalidierung**
   - Kriterien, die auf externen Informationen über die „wahre“ Clusterzugehörigkeiten basieren
3. **Interne Clustervalidierung**
   - Kriterien, die allein auf den Daten basieren

Clustervalidierung | Externe Verfahren

Externe Clustervalidierung (nutzt externe Information)
Es gibt viele Ansätze. Wir diskutieren: *Purity* sowie *Mutual Information*.

Purity (misst die Reinheit von Clustern)

$C_1, \ldots, C_K$ Mengen der Indizes der Datenpunkte jedes Clusters.

$T_1, \ldots, T_K$ Mengen der Indizes der Datenpunkte der wahren Cluster gemäß der externen Information (*ground truth*)

$N_i = |C_i|$ Anzahl Datenpunkte in Cluster *i*

$$\text{purity}_i = \frac{1}{N_i} \max_{j=1}^{K} |C_i \cap T_j|$$

Reinheit (Purity) des Clusters C_i

wird 1, wenn Cluster C_i nur Punkte eines wahren Clusters der Partition *T* enthält

$$\text{purity} = \sum_{i=1}^{K} \frac{N_i}{N} \text{purity}_i = \frac{1}{N} \sum_{i=1}^{K} \max_{j=1}^{K} |C_i \cap T_j|$$

wird 1, wenn alle Cluster C_i nur Punkte eines wahren Clusters enthalten

Clustervalidierung | Externe Verfahren

Purity (misst die Reinheit von Clustern)

$$\text{purity} = \frac{1}{N} \sum_{i=1}^{K} \max_{j=1}^{K} |C_i \cap T_j|$$

Frage F

Mit welchem Vorgehen können Sie *immer* eine Purity von 1 erreichen?

a) Jeder Cluster enthält nur einen Datenpunkt. b) Es gibt nur einen Cluster.

Mutual information (misst Ähnlichkeit zwischen Partitionen C und T)

$$I(C,T) = \sum_{i,j} p_{ij} \log_2 \left(\frac{p_{ij}}{p_{C_i} p_{T_j}} \right)$$

→ kennen Sie bereits aus Vorlesung 5.

Je größer die Mutual Information, desto besser das Clustering.

mit

$$p_{ij} = \frac{|C_i \cap T_j|}{N}$$

Wahrscheinlichkeit, dass ein Punkt aus Cluster i zur Partition j aus T gehört

und

$$p_{C_i} = \frac{|C_i|}{N}$$

$$p_{T_j} = \frac{|T_j|}{N}$$

Wahrscheinlichkeit für Cluster C_i bzw. für Cluster T_j

Clustervalidierung

Interne Clustervalidierung (basiert allein auf vorliegenden Daten)

- In der Praxis am weitesten verbreitet.

Es gibt viele Ansätze. Wir diskutieren:

1. Silhouette / Silhouettenplots
2. Prediction Strength

Clustervalidierung | Interne Verfahren

Silhouette[1]

Wir unterscheiden:

1. **Silhouettenindex**
   misst wie gut ein einzelner Datenpunkt geclustert wurde

   konkret:
   misst die Ähnlichkeit eines Datenpunktes zu den anderen Datenpunkten im Cluster (**Kohäsion**) in Relation zur Unähnlichkeit (**Separation**) mit Datenpunkten anderer Cluster

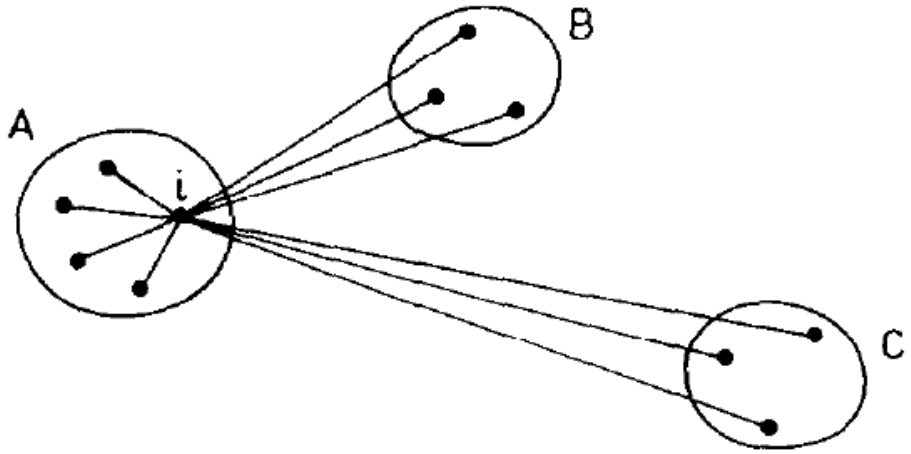
2. **Silhouettenplot**
   visuelle Darstellung der Silhouettenindizes für den gesamten Datensatz, sortiert nach Clusterzugehörigkeit

(1) Rousseeuw, PJ. Journal of Comp. Appl. Math. 20, 53-65 (1987)

Clustervalidierung | Interne Verfahren

Silhouettenindex

Seien C_A, C_B, C_C Mengen, die die Indizes der Datenpunkte in den Clustern A, B und C enthalten.



Sei $i \in C_A$ ein Datenpunkt des Clusters A. Sei d_{ij} die (euklidische) Distanz zwischen Punkten i und j.

$$a(i) := \frac{1}{|C_A|-1} \sum_{j \in C_A, j \neq i} d_{ij}$$

weil wir nicht über d_{ii} summieren

mittlere Distanz zwischen *i* und allen anderen Punkten in Cluster A

→ *Kohäsion*

$$b(i) := \min_{X \neq A} \frac{1}{|C_X|} \sum_{j \in C_X} d_{ij}$$

mittlere Distanz zwischen *i* und allen anderen Punkten im benachbarten Cluster (hier: B)

→ *Separation*

(1) Rousseeuw, PJ. Journal of Comp. Appl. Math. 20, 53-65 (1987)

Clustervalidierung | Interne Verfahren

Silhouettenindex

$$s(i) := \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{falls } a(i) < b(i) \\ 0, & \text{falls } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & \text{falls } a(i) > b(i) \end{cases}$$

$$= \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

i ist gut geclustert:
Unähnlichkeiten im Cluster kleiner als zum benachbarten Cluster

Uneindeutig: sollte *i* zum Cluster A oder B gehören?

i ist falsch geclustert:
liegt im Mittel näher zu Punkten in Cluster B als in Cluster A .

für $|C_A| > 1$, sonst 0.

Wertebereich: $-1 \leq s(i) \leq 1$

Frage **F**

Seien zwei Cluster A und B gegeben. Sei $i \in C_A$ mit Silhouettenindex $s(i)$. Welchen Wert erhalten Sie, wenn Sie *i* dem Cluster B zuordnen?

$s(i) \rightarrow -s(i)$

Clustervalidierung | Interne Verfahren

Mittlere Silhouettenindizes

… für Cluster *X*: $\bar{s}_X = \frac{1}{|C_X|} \sum_{i \in C_X} s(i)$

„average silhouette width“
bewertet einzelne Cluster

… für alle Daten: $\bar{s} = \frac{1}{N} \sum_{i=1}^{N} s(i)$

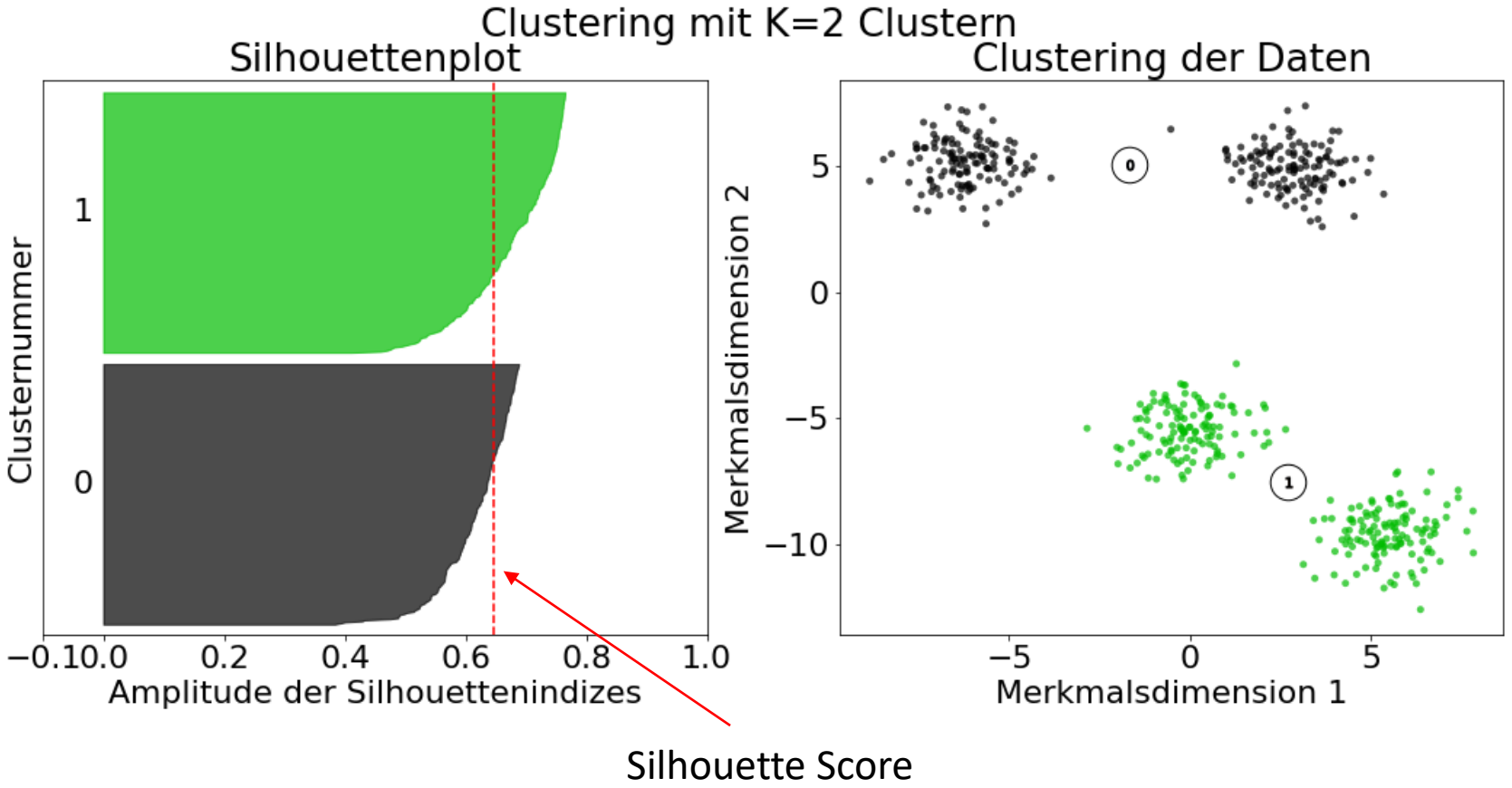
„silhouette score“
bewertet komplettes Clustering

Silhouettenplots

- stellen Silhouettenindizes nach Cluster gruppiert und in absteigender Größe dar.
- erlauben eine visuelle Einschätzung der Clusterqualität (Beispiel auf nachfolgenden Folien)

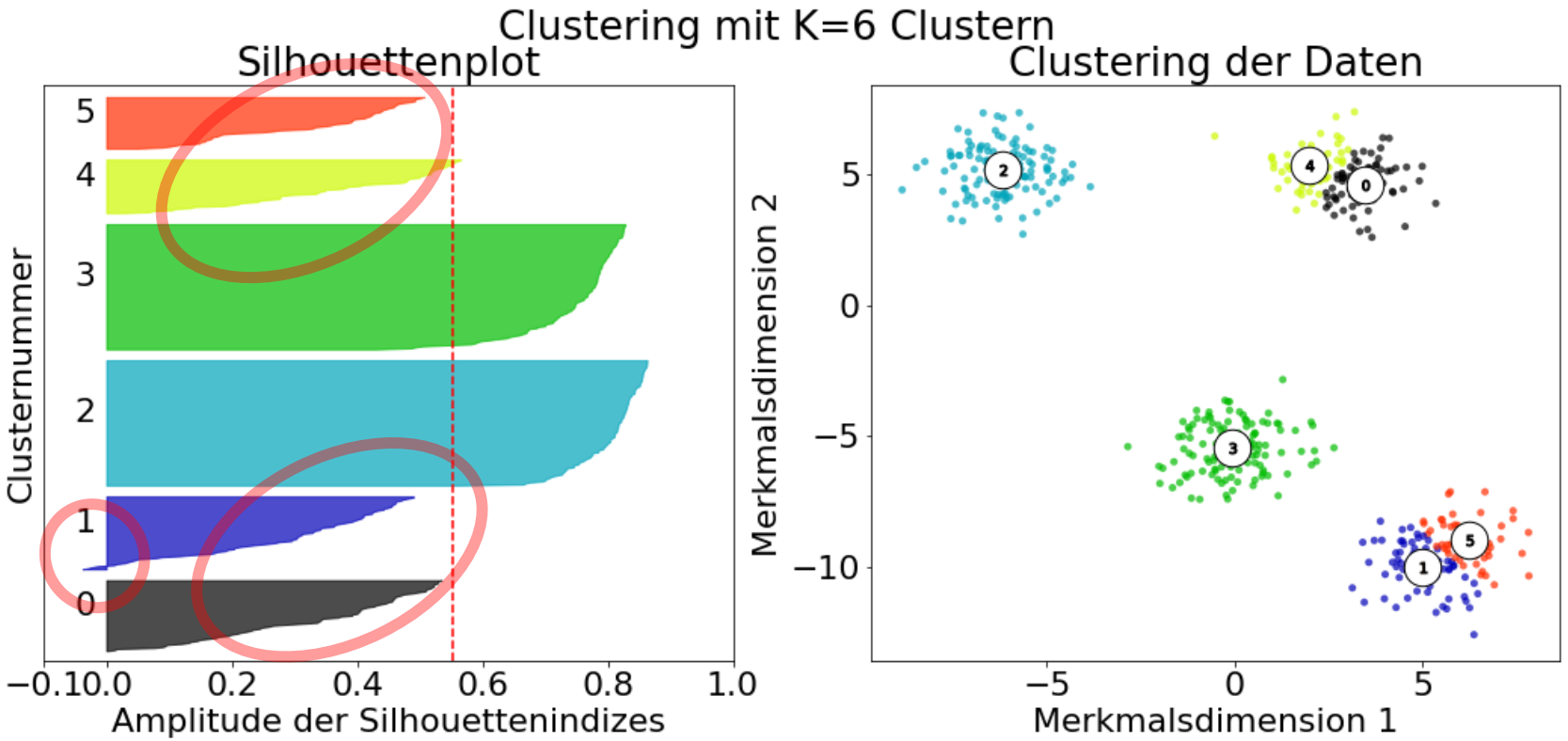
Clustervalidierung | Interne Verfahren

Beispiel: K-Means Clustering eines synthetischen Datensatzes



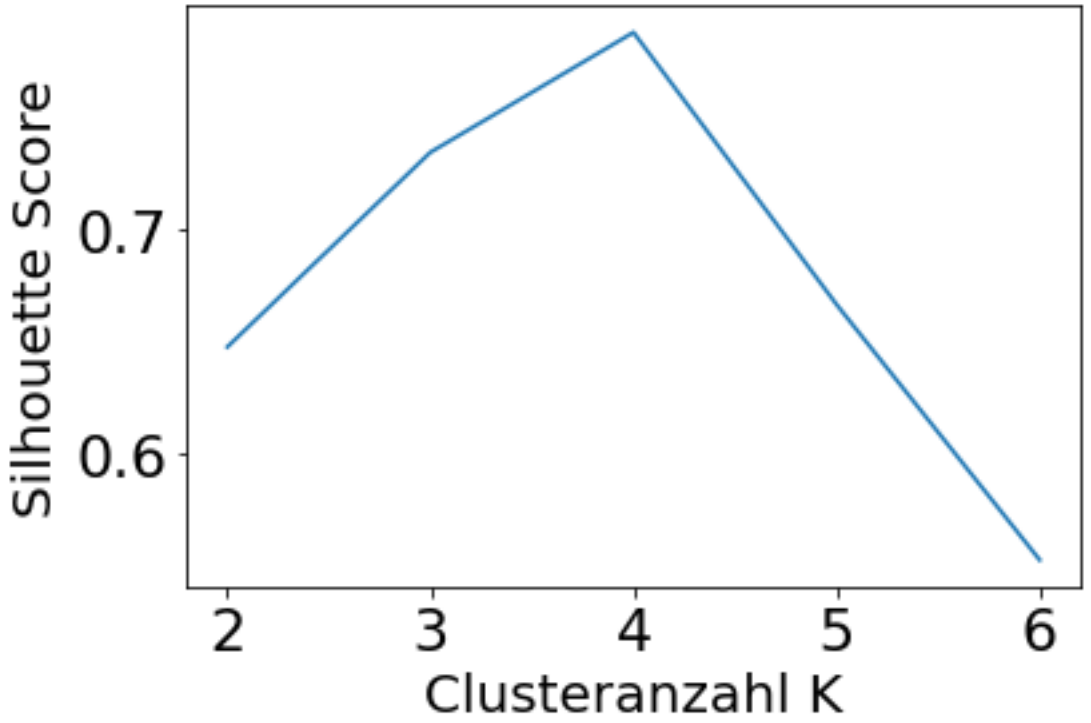
Clustervalidierung | Interne Verfahren

Beispiel: K-Means Clustering eines synthetischen Datensatzes



Clustervalidierung | Interne Verfahren

Beispiel: K-Means Clustering eines synthetischen Datensatzes



Bestimmung der Clusteranzahl über den Silhouetten-Score:
Score zeigt Maximum bei 4 Clustern.

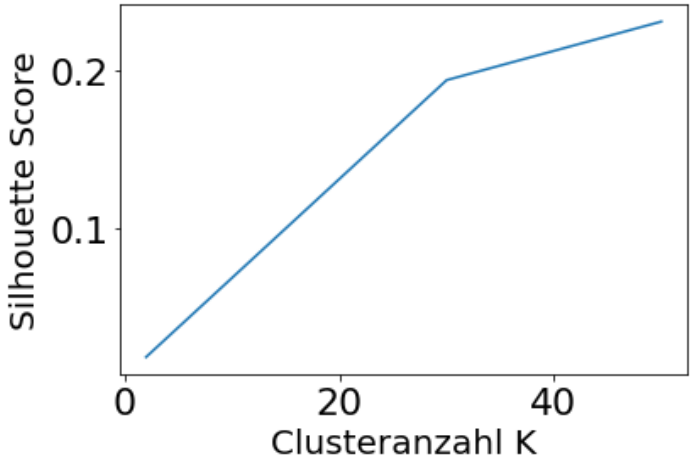
Clustervalidierung | Interne Verfahren

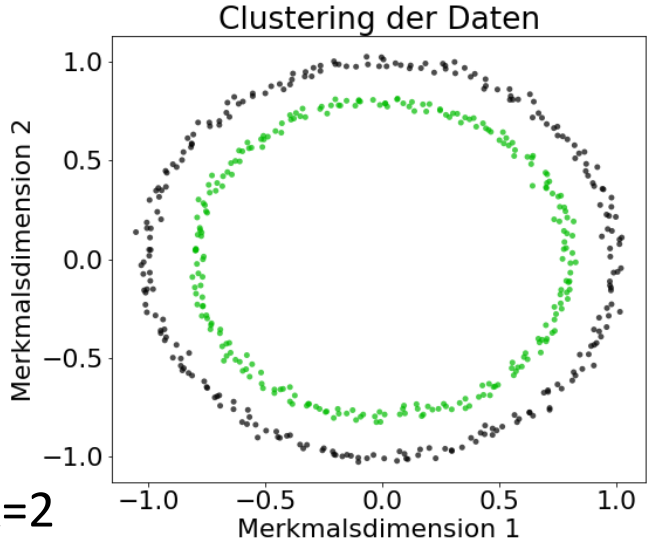
Interaktivität

F

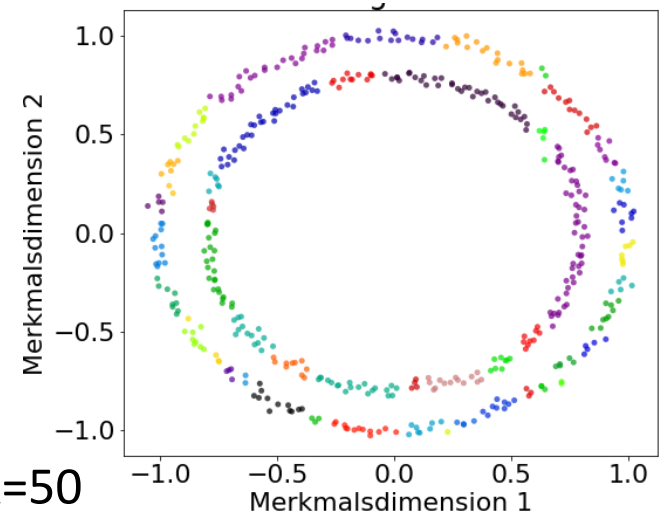
Die Daten (unten) wurden mit *Hierarchical Clustering* (single linkage) geclustert.

Warum zeigt der Silhouetten-Score (rechts) nicht die richtige Clusteranzahl an?





K=2



K=50

Silhouettenplots interpretieren Cluster als kompakte, sphärische Objekte. Dies ist aber nicht notwendigerweise der Fall (wie in diesem Beispiel).

Clustervalidierung

Interne Clustervalidierung (basiert allein auf vorliegenden Daten)

- In der Praxis am weitesten verbreitet.

Es gibt viele Ansätze. Wir diskutieren:

1. Silhouette / Silhouettenplots
2. Prediction Strength

Clustervalidierung | Interne Verfahren

Prediction Strength[1]

Idee: Clusterqualität ist hoch, wenn Clusterzugehörigkeiten auf anderen Realisation der Daten zuverlässig vorhergesagt werden können.

Exkurs (optional; d.h. nur für diejenigen, die sich mit Machine Learning auskennen)

Der *Prediction Strength*-Ansatz interpretiert das Finden richtiger Parameter beim Clustering (wie z.B. Anzahl der Cluster) als *Model Selection* Problem, wie Sie dies aus dem Supervised Machine Learning kennen. Dort werden beste Parameter über die Optimierung der Vorhersagen auf Out-of-Sample Daten (Maximierung der *prediction strength* bzw. Minimierung des Vorhersagefehlers E_{out}) im Rahmen einer Validierung gefunden.

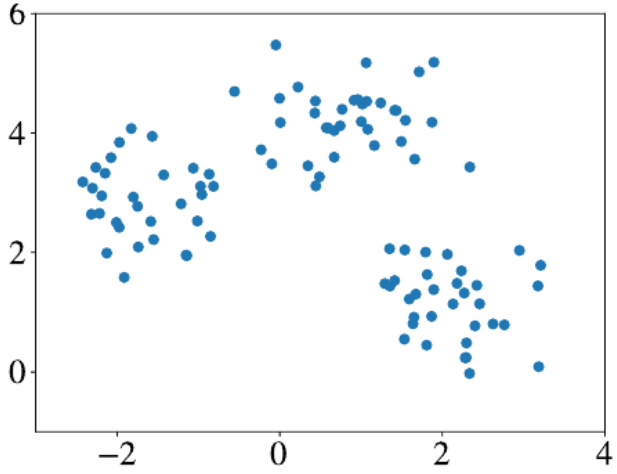
(1) Tibshirani et al. J. Comput. Graph. Stat. 14, 511-528, 2005

Clustervalidierung | Interne Verfahren

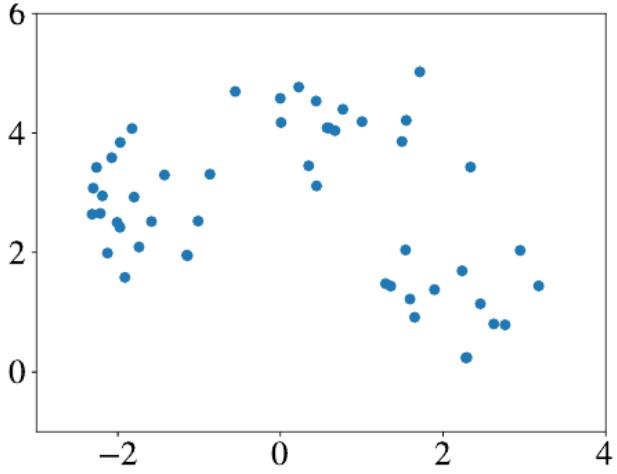
Prediction Strength – Vorgehen

(mathematische Beschreibung folgt später)

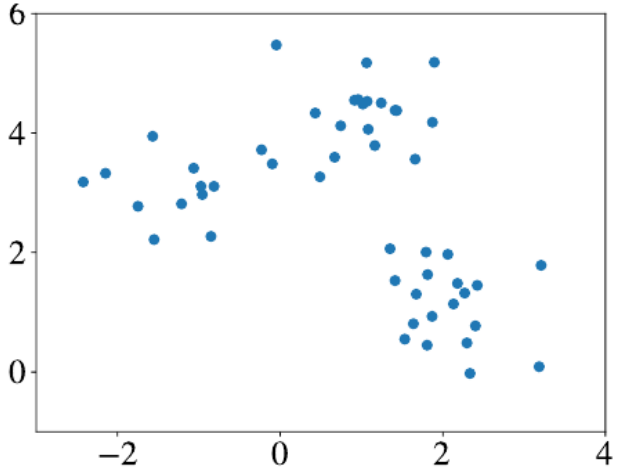
1. Teilen Sie die Daten ($\mathcal{D}$) *zufällig* auf in zwei Mengen: Training- und Validierungsset ($\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$)



$\mathcal{D}$



$\mathcal{D}_{\text{train}}$

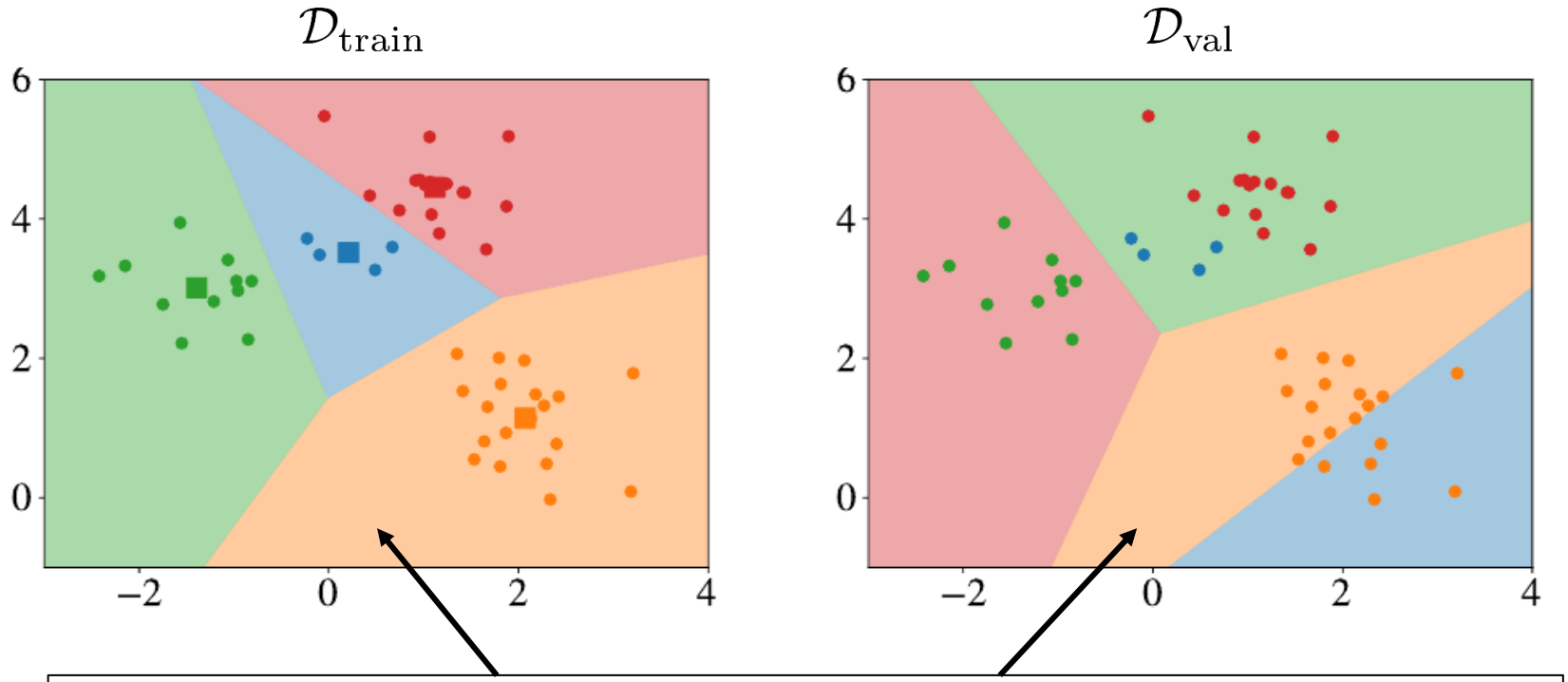


$\mathcal{D}_{\text{val}}$

Clustervalidierung | Interne Verfahren

2. Clustern Sie die Trainingsdaten und Validierungsdaten separat mit denselben gewählten Parametern (z.B. Clusteranzahl K)

Beispiel: Cluster ermittelt über K-Means mit $K = 4$

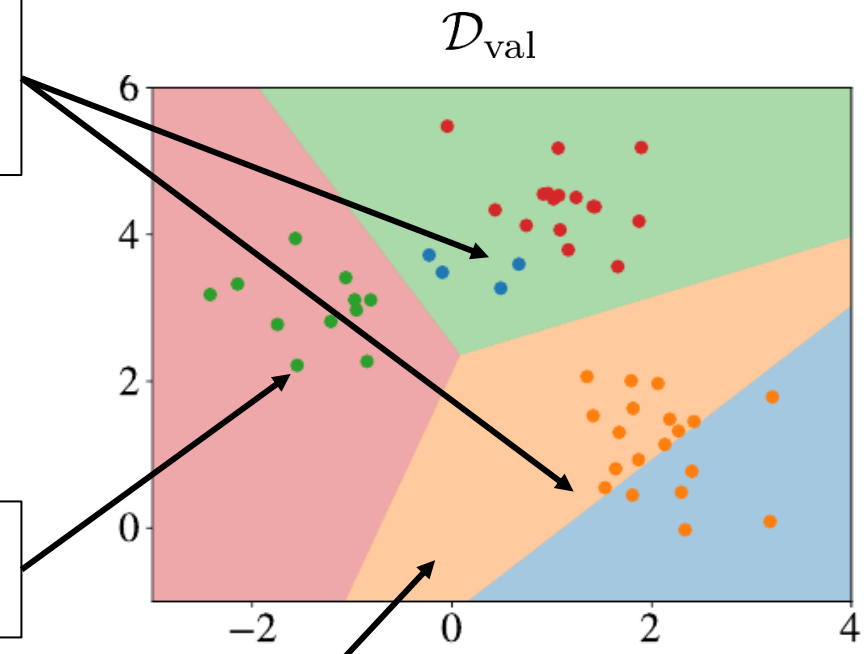


Farbige Flächen: Cluster gefunden über K-Means in den jeweiligen Datensätzen

Clustervalidierung | Interne Verfahren

3. Ermitteln Sie für alle Daten in $\mathcal{D}_{\text{val}}$ die Clusterzugehörigkeit gemäß des Clustering in $\mathcal{D}_{\text{train}}$.

Beobachtung
Einige Punkte in $\mathcal{D}_{\text{val}}$ werden in andere Cluster aufgeteilt als in $\mathcal{D}_{\text{train}}$



Farbige Punkte: Cluster-zugehörigkeit gemäß $\mathcal{D}_{\text{train}}$

Farbige Flächen: Cluster gefunden über K-Means in $\mathcal{D}_{\text{val}}$

Clustervalidierung | Interne Verfahren

4. Bestimmen Sie für jeden Cluster in $\mathcal{D}_{\text{val}}$ den Bruchteil *p* aller Paare von Datenpunkten, die sich auch im selben Cluster in $\mathcal{D}_{\text{train}}$ befinden würden.

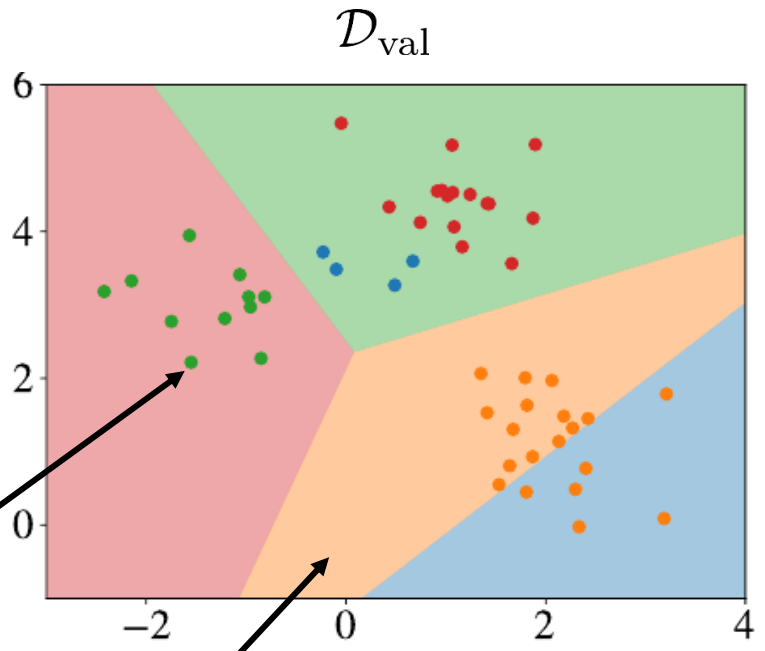
Der kleinste Wert *p* über aller Cluster heißt *prediction strength*.

F

Frage
Welchen Wert *p* hat der
- rote Cluster in $\mathcal{D}_{\text{val}}$? $p = 1$
- blaue Cluster in $\mathcal{D}_{\text{val}}$? $p = 1$
- grüne Cluster in $\mathcal{D}_{\text{val}}$?

$$N_{\text{Paare}} = \frac{19 \cdot 18}{2}$$

$$p = \frac{N_{\text{Paare}} - 4 \cdot 15}{N_{\text{Paare}}} \approx 0.65$$



Farbige Punkte: Cluster-zugehörigkeit gemäß $\mathcal{D}_{\text{train}}$

Farbige Flächen: Cluster gefunden über K-Means in $\mathcal{D}_{\text{val}}$

Clustervalidierung | Interne Verfahren

Prediction Strength – formale Behandlung

Sei $\mathcal{C} = \{C_1, \ldots, C_K\}$ ein Clustering von $\mathcal{D}_{\text{train}}$ mit Indexmengen C_i (Cluster in $\mathcal{D}_{\text{train}}$), deren Datenpunkte in den jeweiligen Regionen R_{C_i} des Merkmalsraum liegen.

Sei $\mathcal{A} = \{A_1, \ldots, A_K\}$ ein Clustering von $\mathcal{D}_{\text{val}}$ mit Indexmengen $A_1, \ldots, A_K$ (Cluster in $\mathcal{D}_{\text{val}}$).

Sei M eine $N_{\text{val}} \times N_{\text{val}}$ Matrix (sog. *Ko-Mitgliedschaftsmatrix*), wobei N_{val} die Anzahl der Datenpunkte in $\mathcal{D}_{\text{val}}$ bezeichnet:

$$M_{ii'} := \begin{cases} 1, & \exists k \text{ mit } (i, i') \in R_{C_k} \\ 0, & \text{sonst} \end{cases}$$

d.h.: Matrixeintrag ist 1, wenn die jeweiligen zwei Datenpunkte aus dem **Validierungsset** zur selben Region eines Clusters *k* im **Trainingset** gehören

Prediction strength:

$$\text{ps}(K) = \min_{j=1,\ldots,K} \left[\frac{1}{|A_j|(|A_j|-1)} \sum_{i,i' \in A_j, i \neq i'} M_{ii'} \right]$$

Summe über Paare im Cluster A_j des Validierungssets

Clustervalidierung | Interne Verfahren

Wie werden die Regionen R_{C_i} ermittelt?

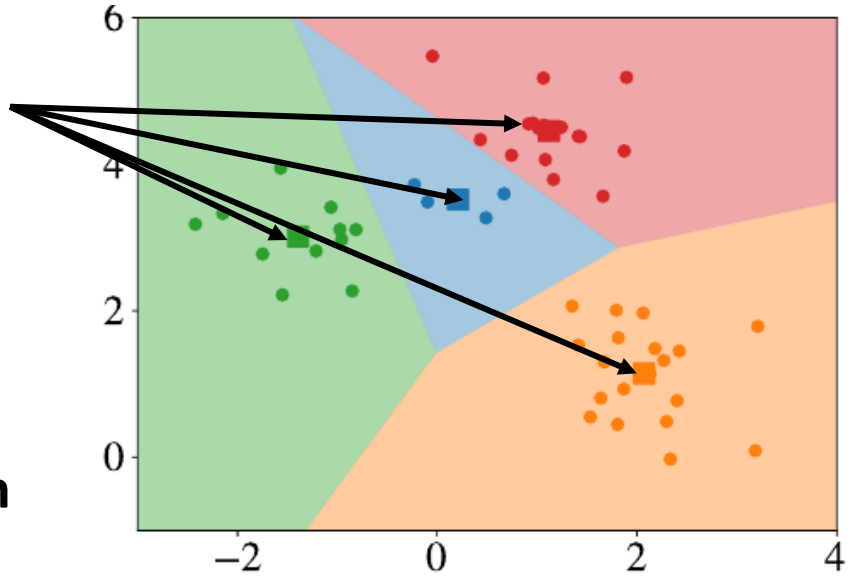
Ermittlung der Regionen

- **hängt vom jeweiligen Clusterverfahren ab**
- kodiert Vorstellung davon, was ein Cluster ist

Beispiel 1

K-Means

- interpretiert Cluster als sphärische Objekte im Merkmalsraum
- Regionen sind Polygone:
  Centroids der Cluster C_i definieren Regionen und damit Clusterzugehörigkeit!



Clustervalidierung | Interne Verfahren

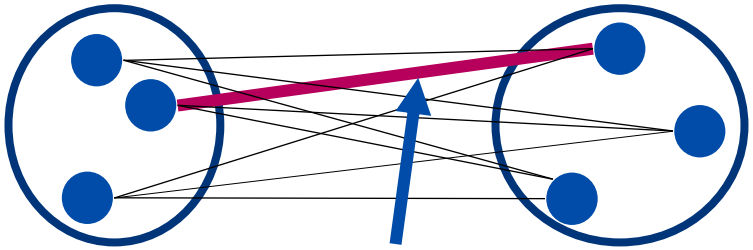
Wie werden die Regionen R_{C_i} ermittelt?

Beispiel 2

Single Linkage Clustering (Vorlesung 8)

- betrachtete Abstände sind kleinste euklidische Distanzen
- Datenpunkt *i* aus $\mathcal{D}_{\text{val}}$ wird dem Cluster aus $\mathcal{D}_{\text{train}}$ zugeordnet, zu dem er den kleinsten euklidischen Abstand hat
  1. Für gegebenen Datenpunkt *i* aus $\mathcal{D}_{\text{val}}$ **suche nächsten Nachbarn** *j* in $\mathcal{D}_{\text{train}}$
  2. Übernehme Clustermitgliedschaft von *j* für den Punkt *i*

Unterschiedlichkeitsmaß bei Single Linkage Clustering:



Minimale euklidische Distanz entspricht Unterschiedlichkeit beider Cluster

Clustervalidierung | Interne Verfahren

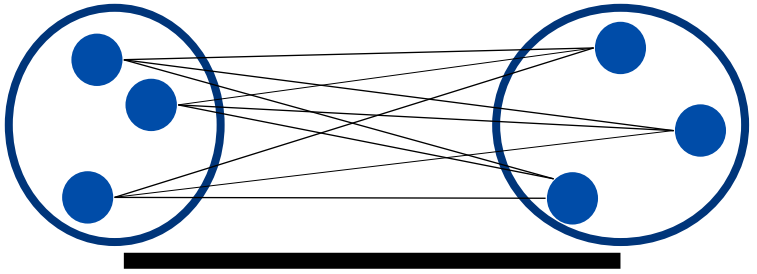
Wie werden die Regionen R_{C_i} ermittelt?

Beispiel 3

Average Linkage Clustering (Vorlesung 8)

- betrachtete Abstände sind mittlere euklidische Distanzen
- Weise Datenpunkt *i* aus $\mathcal{D}_{\text{val}}$ dem Cluster aus $\mathcal{D}_{\text{train}}$ zu, zu dem er den kleinsten mittleren euklidischen Abstand hat.

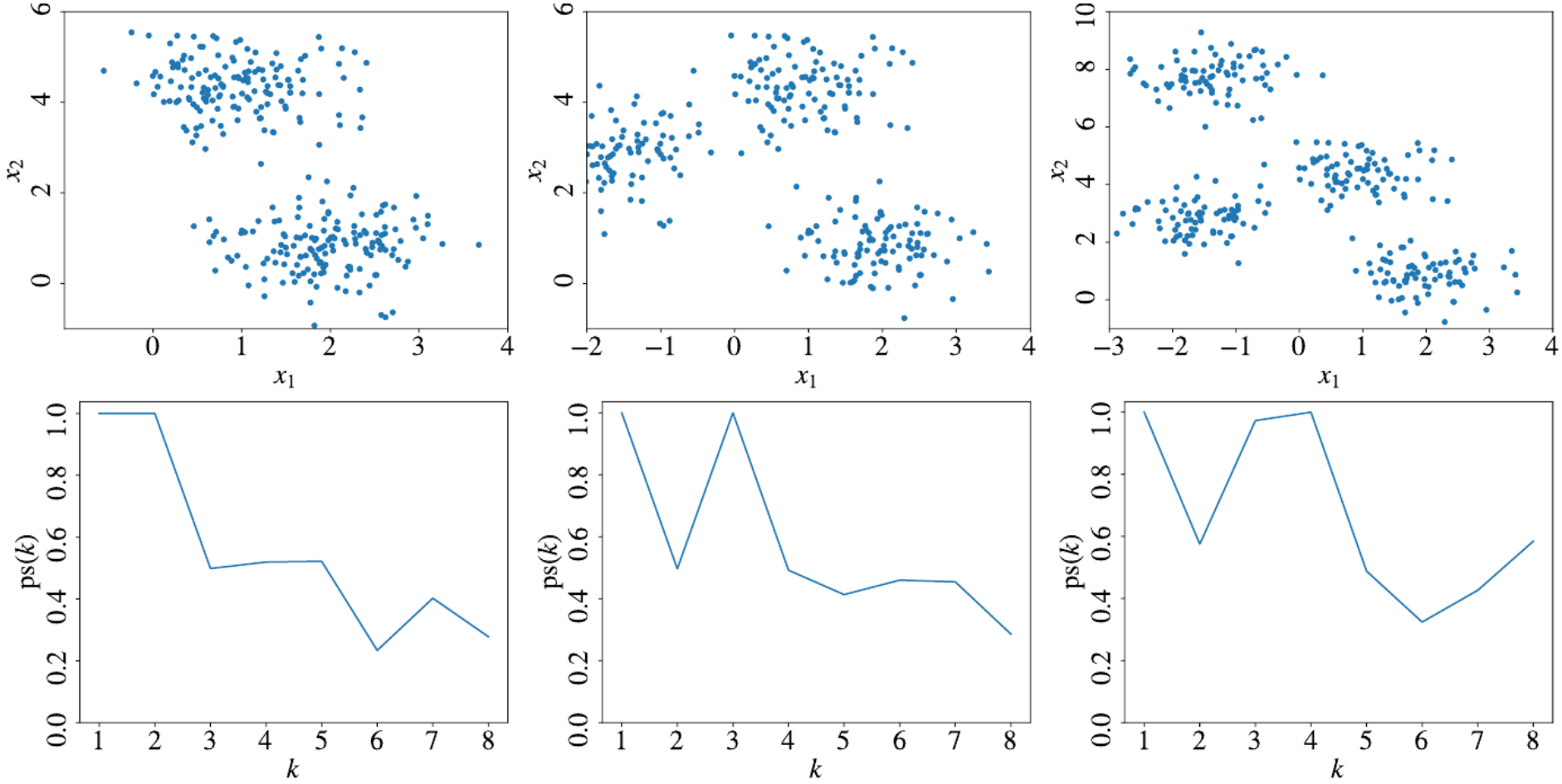
(aus Vorlesung 8)



mittlerer euklidischer Abstand

Clustervalidierung | Interne Verfahren

Beispiel – K-Means Clustering

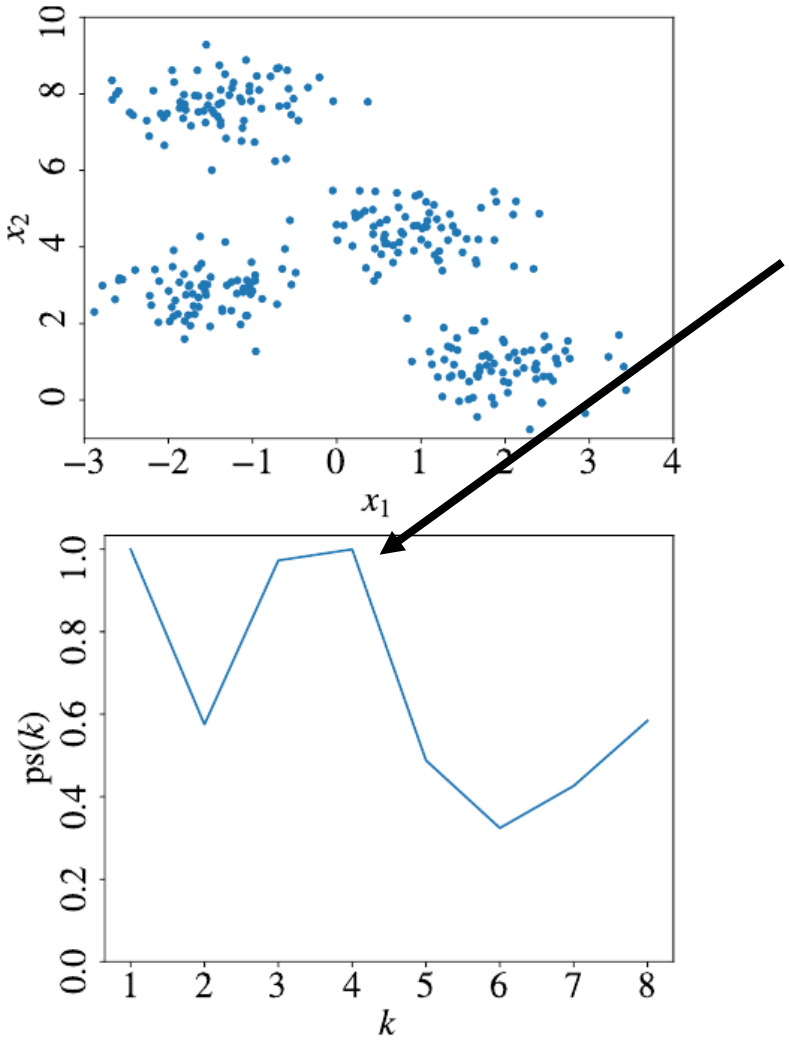


Frage: Warum ist ps(1) = 1?

Training- und Testset-Datenpunkte sind in nur einem Cluster enthalten.

F

Clustervalidierung | Interne Verfahren



Hinweise

- Optimale Clusteranzahl *k* entspricht größter Prediction Strength für größten Wert von k
- (Häufig) Datenaufteilung in 80% Training-, 20% Validierungsset
- für die Machine Learning Spezialisten:
  Bestimmung der Prediction Strength per Kreuzvalidierung (je nach Rechenaufwand)