

Mathematische Grundlagen III

Maschinelles Lernen III: Clustering

Prof. Dr. Matthew Crocker

Universität des Saarlandes

13. Juli 2015

Clustering vs. Klassifikation

In den letzten beiden Vorlesungen haben wir uns mit Klassifikationsalgorithmen (Naive Bayes Klassifizierer und Entscheidungsbäumen) beschäftigt.

Heute schauen wir uns Methoden für **Clustering** an.

Unterschiede Klassifikation und Clustering

- Bei der Klassifikation werden Instanzen vordefinierten Klassen zugeordnet → **supervised**
- Beim Clustering entdeckt der Algorithmus “natürliche” Klassen, die die Instanzen in Gruppen mit ähnlichen Eigenschaften teilen → **unsupervised**
- Deutscher Begriff: *Ballungsanalyse*

Wozu Clustering verwenden?

Explorative Datenanalyse

Um ein Gefühl für die vorhandenen Daten und ihre Eigenschaften zu gewinnen

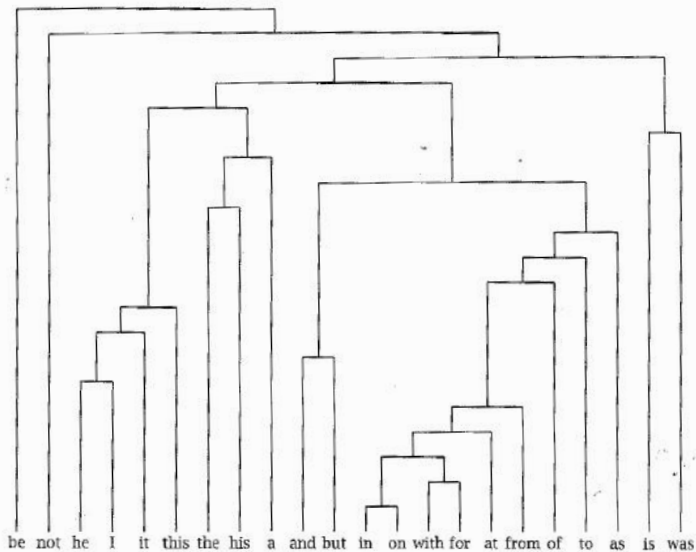
Binning

Instanzen entdecken, die sich ähnlich verhalten und daher ähnlich behandelt werden können, um Abhilfe bei Sparse-Data-Problemen zu schaffen

Beispiel:

- In einem Korpus findet man die Sequenzen “*am Donnerstag*” und “*am Freitag*” sowie “*donnertags*” und “*freitags*”
- Außerdem hat man “*am Montag*”, aber *montags* kommt nicht vor
- Wenn wir wissen, dass *Donnerstag*, *Freitag* und *Montag* sich syntaktisch ähnlich verhalten, können wir *montags* inferieren

Beispiel: häufige englische Wörter



Inhaltsverzeichnis

- 1 Verschiedene Arten von Clustering
- 2 Clusteringmethoden und -algorithmen
 - Hierarchisches Clustering
 - Flaches Clustering
- 3 Anwendungsbeispiel: Nach semantischer Ähnlichkeit clustern
- 4 Evaluation von Clustering Modellen
- 5 Course Wrap-up

Inhaltsverzeichnis

- 1 **Verschiedene Arten von Clustering**
- 2 Clusteringmethoden und -algorithmen
 - Hierarchisches Clustering
 - Flaches Clustering
- 3 Anwendungsbeispiel: Nach semantischer Ähnlichkeit clustern
- 4 Evaluation von Clustering Modellen
- 5 Course Wrap-up

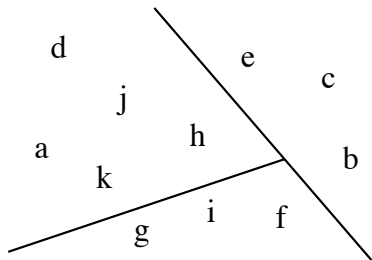
Arten von Clustering

Clusteringalgorithmen können verschiedene Arten von Clustern erzeugen:

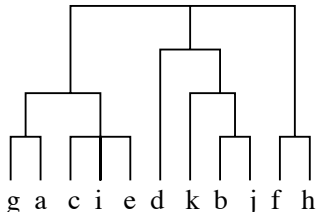
- **Iterativ:** Der Algorithmus beginnt mit einer Anfangsmenge von Clustern und verbessert diese immer weiter.
- **Hierarchisch oder flach:** hierarchische Algorithmen generieren eine Hierarchie von Clustern, sodass es verschiedene Granularitäten gibt. Bei flachen Algorithmen gibt es nur ein Granularitätslevel und alle Cluster sind gleichwertig.
- **Disjunktiv:** Eine Instanz kann mehreren Clustern zugeordnet werden.
- **Hart oder weich:** Bei hartem Clustering wird jede Instanz genau einem Cluster zugeordnet, bei weichem Clustering wird eine Instanz einem Cluster mit bestimmter Wahrscheinlichkeit zugeordnet.

Verschiedene Clustering-Arten

Flach



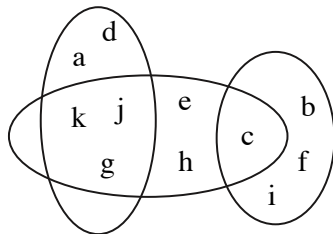
Hierarchisch (Dendrogramme)



Verschiedene Clustering-Arten

Disjunktiv

Instanzen können mehreren Cluster angehören



Probabilistisch oder "soft"

Für jeden Cluster wird eine Wahrscheinlichkeit angegeben, dass eine Instanz ihm zugeordnet wird

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

Inhaltsverzeichnis

- 1 Verschiedene Arten von Clustering
- 2 Clusteringmethoden und -algorithmen
 - Hierarchisches Clustering
 - Flaches Clustering
- 3 Anwendungsbeispiel: Nach semantischer Ähnlichkeit clustern
- 4 Evaluation von Clustering Modellen
- 5 Course Wrap-up

Eigenschaften von Clusteringalgorithmen

Hierarchisches Clustering

- gut für detaillierte Datenanalyse
- mehr Information als flaches Clustering
- bester Algorithmus hängt von der Anwendung ab
- weniger effizient als flaches Clustering

Flaches Clustering

- gut, wenn Effizienz wichtig ist
- **k-means Clustering** ist ein einfacher Algorithmus dafür, Resultate oft ausreichend.
- Voraussetzung für k-means clustering: Daten können in Euklidischen Raum dargestellt werden.
- Alternativ: EM Algorithmus

Inhaltsverzeichnis

- 1 Verschiedene Arten von Clustering
- 2 Clusteringmethoden und -algorithmen
 - Hierarchisches Clustering
 - Flaches Clustering
- 3 Anwendungsbeispiel: Nach semantischer Ähnlichkeit clustern
- 4 Evaluation von Clustering Modellen
- 5 Course Wrap-up

Hierarchisches Clustering

Beim hierarchischen Clustering wollen wir einen Baum generieren, der beschreibt, wie stark sich die verschiedenen Instanzen / Gruppen von Instanzen ähneln.

- **Bottom-up** agglomeratives Clustering
Jede Instanz ist ein Cluster. Gruppiere die zwei ähnlichsten Cluster zu einem neuen Cluster.
- **Top-down** divisives Clustering
fange an mit Cluster, das alle Instanzen enthält, und teile das am wenigsten kohärente Cluster in zwei Cluster auf.

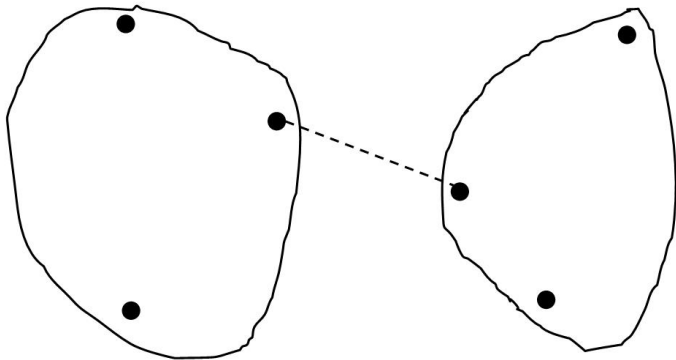
Berechnung der Ähnlichkeit

Single Link: Distanz der ähnlichsten Instanzen zweier Cluster

Complete Link: Distanz der entferntesten Instanzen zweier Cluster

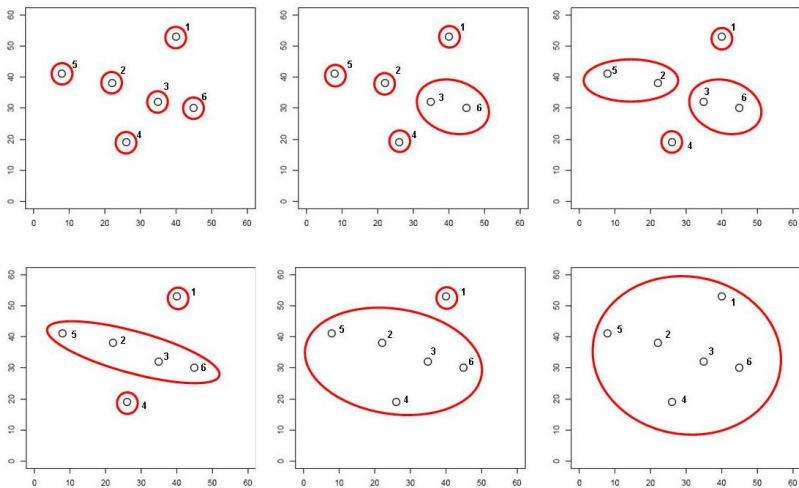
Group-Average: durchschnittliche Distanz der Instanzen zweier Cluster

Single Link Clustering: Beispiel

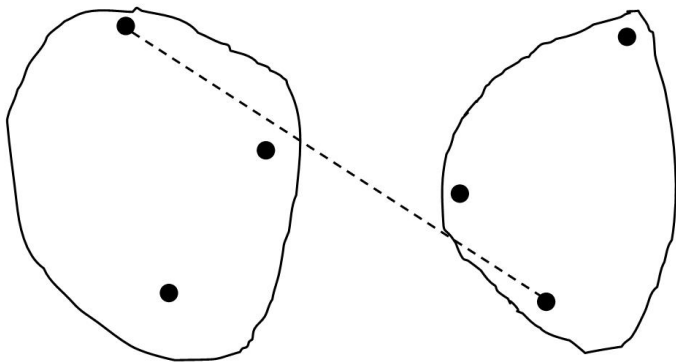


Graphik von Clustergruppe TU München

Single Link Clustering: Beispiel

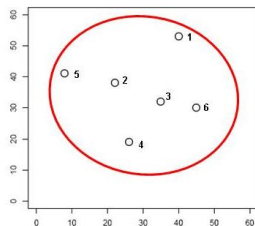
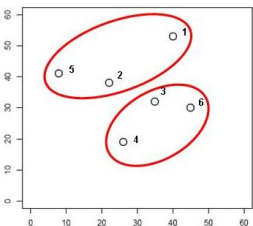
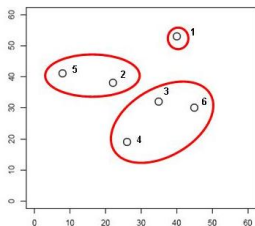
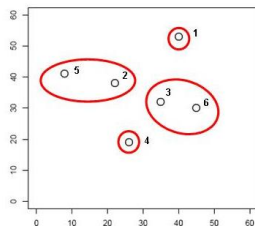
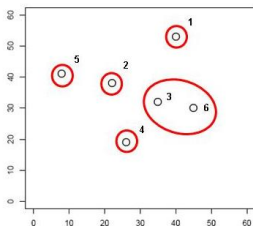
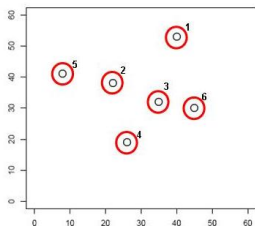


Complete Link Clustering: Beispiel

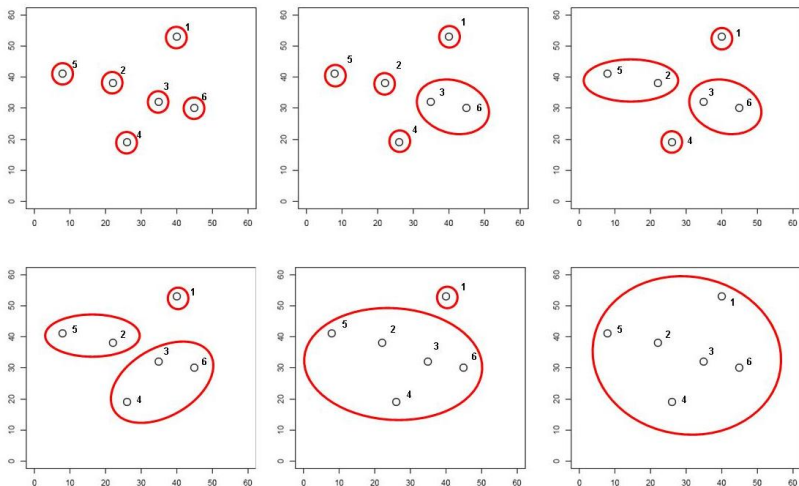


Graphik von Clustergruppe TU München

Complete Link Clustering: Beispiel



Group Average Clustering: Beispiel



Ähnlichkeitsfunktionen beim Clustering

- single link, complete link, group-average
- single link führt zu elongierten Clustern
- complete link verhindert dies
- group-average beschreibt ein “rundes” Clustering um einen Mittelpunkt.
- Welche Funktion angebracht ist, hängt von den Daten / der Anwendung ab.

Inhaltsverzeichnis

- 1 Verschiedene Arten von Clustering
- 2 Clusteringmethoden und -algorithmen
 - Hierarchisches Clustering
 - Flaches Clustering
- 3 Anwendungsbeispiel: Nach semantischer Ähnlichkeit clustern
- 4 Evaluation von Clustering Modellen
- 5 Course Wrap-up

Iteratives distanz-basiertes Clustering

Intuition bei *k-means*

- Bestimme k , die Anzahl der gewünschten Cluster
- Wähle k beliebige Punkte als Cluster-Zentren aus
- Weise jede Instanz dem nächsten Cluster-Zentrum zu
- Berechne den Mittelpunkt für jeden Cluster und verwende ihn als neues Zentrum
- Weise alle Instanzen wieder dem nächsten Cluster-Zentrum zu
- Iteriere, bis alle Cluster stabil sind

Iteratives distanz-basiertes Clustering

Der Algorithmus

- Jede Instanz \vec{x} im Training Set wird als Vektor mit einem Wert pro Attribut repräsentiert

$$\vec{x} = (x_1, x_2, \dots, x_n)$$

- Die Distanz zwischen zwei Vektors \vec{x} and \vec{y} ist definiert als (euklidische Distanz):

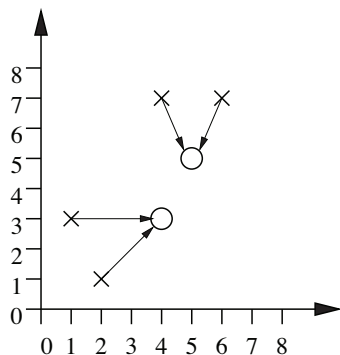
$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Der Mittelpunkt $\vec{\mu}$ einer Menge Vektoren c_j ist definiert als:

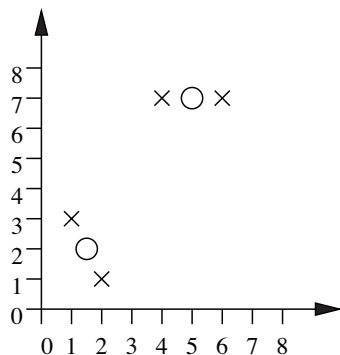
$$\vec{\mu} = \frac{1}{|c_j|} \sum_{\vec{x} \in c_j} \vec{x}$$

Iteratives distanz-basiertes Clustering

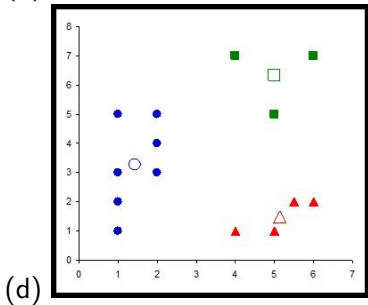
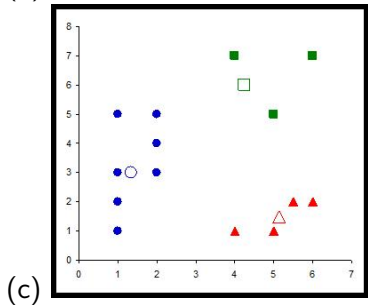
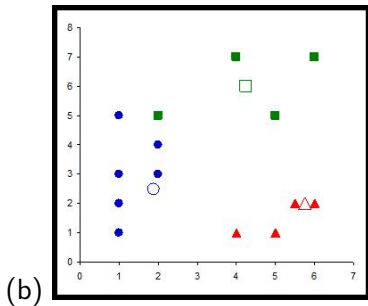
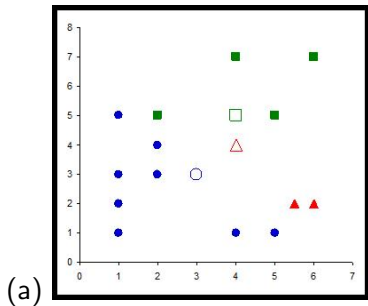
Beispiel



Die Instanzen (Kreuzchen) werden anfangs zum nächsten Cluster-Zentrum (Kreise) zugewiesen



Der Mittelpunkt jedes Cluster wird dann berechnet und als neuen Zentrum verwendet



Weiteres Beispiel entnommen: Clustergruppe TU München

Iteratives distanz-basiertes Clustering

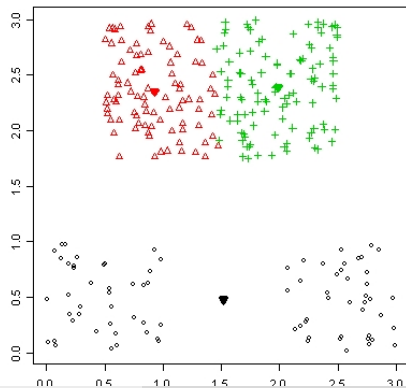
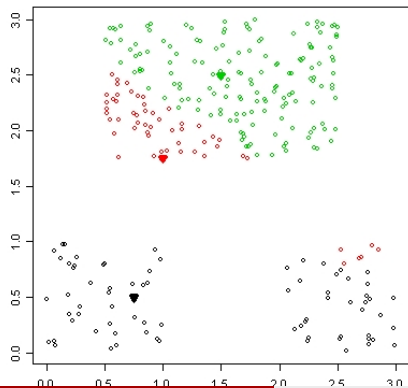
Eigenschaften von k-means

- Flaches Clustering-Verfahren
- Effizient bei großen Datenmengen
- Nicht geeignet für Nominaldaten
- Findet nur ein lokales Maximum, keine globales
- Die Cluster hängen stark ab von der initialen Wahl der Clusterzentren
- weiche Version von k-means: EM Algorithmus

Iteratives distanz-basiertes Clustering

Eigenschaften von k-means

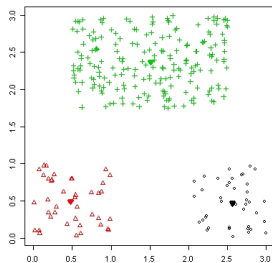
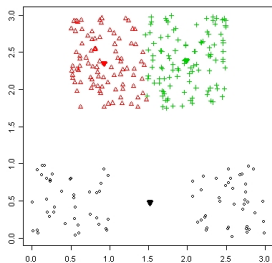
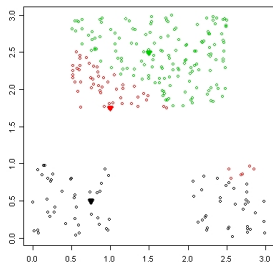
- Findet nur ein lokales Maximum, keine globales
- Die Cluster hängen stark ab von der initialen Wahl der Clusterzentren
- weiche Version von k-means: EM Algorithmus



Iteratives distanz-basiertes Clustering

Eigenschaften von k-means

- Flaches Clustering-Verfahren
- Effizient bei großen Datenmengen
- Nicht geeignet für Nominaldaten
- Findet nur ein lokales Maximum, keine globales
- Die Cluster hängen stark ab von der initialen Wahl der Clusterzentren
- weiche Version von k-means: EM Algorithmus



Iteratives distanz-basiertes Clustering

Eigenschaften von k-means

- Flaches Clustering-Verfahren
- Effizient bei großen Datenmengen
- Nicht geeignet für Nominaldaten
- Findet nur ein lokales Maximum, keine globales
- Die Cluster hängen stark ab von der initialen Wahl der Clusterzentren
- weiche Version von k-means: EM Algorithmus
- Kann man für hierarchisches Clustering verwenden
erst k-means mit $k=2$ anwenden, dann nochmal für jedes der beiden Cluster anwenden und so weiter.

Inhaltsverzeichnis

- 1 Verschiedene Arten von Clustering
- 2 Clusteringmethoden und -algorithmen
 - Hierarchisches Clustering
 - Flaches Clustering
- 3 Anwendungsbeispiel: Nach semantischer Ähnlichkeit clustern**
- 4 Evaluation von Clustering Modellen
- 5 Course Wrap-up

Anwendung: Semantisch ähnliche Wörter finden

Warum kann es uns helfen wenn wir wissen welche Wörter semantisch ähnlich sind?

- Probleme mit Datenspärlichkeit umgehen, indem wir Wahrscheinlichkeiten über ähnliche Ereignisse abschätzen

Beispiel

Wir wollen die Wörter "Strand", "Meer", "Student" und "Klausur" clustern.

Nehmen wir an, wir beobachten folgende Kookkurrenzen mit den Wörtern "Sand", "Sonne", "Uni" und "lernen".

	Strand	Meer	Student	Klausur
Sand	12	11	0	1
Sonne	7	5	0	2
Uni	2	0	9	15
lernen	0	1	14	10

Clustering Beispiele mit hierarchical clustering und k-means

Zu clustern: "Strand", "Meer", "Student" und "Klausur"

	Strand	Meer	Student	Klausur
Sand	12	11	0	1
Sonne	7	5	0	2
Uni	2	0	9	15
lernen	0	1	14	10

Formel : $|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$ wobei : $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$

Rechnung: Distanz zwischen "Strand" und "Meer"

$$\vec{Strand} = \begin{pmatrix} 12 \\ 7 \\ 2 \\ 0 \end{pmatrix} \text{ und } \vec{Meer} = \begin{pmatrix} 11 \\ 5 \\ 0 \\ 1 \end{pmatrix}$$

$$|\vec{Strand} - \vec{Meer}| = ?$$

Clustering Beispiele mit hierarchical clustering und k-means

Zu clustern: "Strand", "Meer", "Student" und "Klausur"

	Strand	Meer	Student	Klausur
Sand	12	11	0	1
Sonne	7	5	0	2
Uni	2	0	9	15
lernen	0	1	14	10

$$\text{Formel: } |\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad \text{wobei: } \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

Rechnung: Distanz zwischen "Strand" und "Meer"

$$\vec{\text{Strand}} = \begin{pmatrix} 12 \\ 7 \\ 2 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{\text{Meer}} = \begin{pmatrix} 11 \\ 5 \\ 0 \\ 1 \end{pmatrix}$$

$$\begin{aligned} |\vec{\text{Strand}} - \vec{\text{Meer}}| &= \sqrt{(12 - 11)^2 + (7 - 5)^2 + (2 - 0)^2 + (0 - 1)^2} \\ &= \sqrt{1 + 4 + 4 + 1} = \sqrt{10} = 3.16 \end{aligned}$$

Clustering Beispiele mit hierarchical clustering und k-means

Zu clustern: "Strand", "Meer", "Student" und "Klausur"

	Strand	Meer	Student	Klausur
Sand	12	11	0	1
Sonne	7	5	0	2
Uni	2	0	9	15
lernen	0	1	14	10

$$\text{Formel: } |\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad \text{wobei: } \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

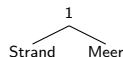
Rechnung: Distanz zwischen "Strand" und "Student":

$$\vec{\text{Strand}} = \begin{pmatrix} 12 \\ 7 \\ 2 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{\text{Student}} = \begin{pmatrix} 0 \\ 0 \\ 9 \\ 14 \end{pmatrix}$$

$$\begin{aligned} |\vec{\text{Strand}} - \vec{\text{Student}}| &= \sqrt{(12 - 0)^2 + (7 - 0)^2 + (2 - 9)^2 + (0 - 14)^2} \\ &= \sqrt{144 + 49 + 49 + 196} = \sqrt{438} = 20.92 \end{aligned}$$

Beispiel

Distanzen	Strand	Meer	Student	Klausur
Strand	0	3.16	20.92	20.37
Meer	–	0	19.89	20.37
Student	–	–	0	7.55
Klausur	–	–	–	0

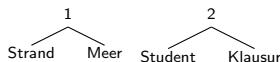


Beispiel 1: Single Link Agglomerative Clustering

- ① Jede Instanz ist ein Cluster
- ② Finde die beiden Cluster mit kleinster Distanz: Strand – Meer.
- ③ Fasse diese zu einem Cluster zusammen.

Beispiel

Distanzen	Strand	Meer	Student	Klausur
Strand	0	3.16	20.92	20.37
Meer	–	0	19.89	20.37
Student	–	–	0	7.55
Klausur	–	–	–	0

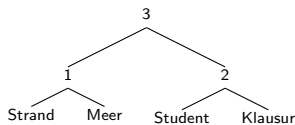


Beispiel 1: Single Link Agglomerative Clustering

- ① Jede Instanz ist ein Cluster
- ② Finde die beiden Cluster mit kleinster Distanz: Strand – Meer.
- ③ Fasse diese zu einem Cluster zusammen.
- ④ Fasse die beiden Cluster zusammen, die einander ähnlichsten Elemente haben: Student – Klausur.

Beispiel

Distanzen	Strand	Meer	Student	Klausur
Strand	0	3.16	20.92	20.37
Meer	–	0	19.89	20.37
Student	–	–	0	7.55
Klausur	–	–	–	0



Beispiel 1: Single Link Agglomerative Clustering

- ① Jede Instanz ist ein Cluster
- ② Finde die beiden Cluster mit kleinster Distanz: Strand – Meer.
- ③ Fasse diese zu einem Cluster zusammen.
- ④ Fasse die beiden Cluster zusammen, die einander ähnlichsten Elemente haben: Student – Klausur.
- ⑤ Fasse die beiden Cluster zusammen, die einander ähnlichsten Elemente haben: Meer – Student.

Beispiel

Distanzen	Strand	Meer	Student	Klausur	Z1	Z2
Strand	0	3.16	20.92	20.37		
Meer	-	0	19.89	20.37		
Student	-	-	0	7.55		
Klausur	-	-	-	0		

Beispiel 2: k-means Clustering

1 Lege 2 zufällige Clusterzentren fest: $\vec{z}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, $\vec{z}_2 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$

2 Berechne Distanzen zu Instanzen:

$$|\vec{Strand} - \vec{z}_1| = \sqrt{(12-1)^2 + (7-1)^2 + (2-1)^2 + (0-1)^2} = 12.60$$

3 Ordne die Instanzen den nächsten Clusterzentren zu

4 Berechne Werte für $Z1_{neu}$ und $Z2_{neu}$ mit der Formel $\vec{\mu} = \frac{1}{|c_j|} \sum_{\vec{x} \in c_j} \vec{x}$

$$Z1_{neu}^{\vec{}} = \frac{1}{2}(\vec{Strand} + \vec{Meer}) = \frac{1}{2} \begin{pmatrix} 12 \\ 7 \\ 2 \\ 0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 11 \\ 5 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 11.5 \\ 6 \\ 1 \\ 0.5 \end{pmatrix}$$

Beispiel

Distanzen	Strand	Meer	Student	Klausur	Z1	Z2
Strand	0	3.16	20.92	20.37	12.60	12.76
Meer	-	0	19.89	20.37	10.81	11.26
Student	-	-	0	7.55	15.32	11.87
Klausur	-	-	-	0	16.67	13.42

Beispiel 2: k-means Clustering

① Lege 2 zufällige Clusterzentren fest: $\vec{z}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, $\vec{z}_2 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$

② Berechne Distanzen zu Instanzen:

$$|\vec{Strand} - \vec{z}_1| = \sqrt{(12-1)^2 + (7-1)^2 + (2-1)^2 + (0-1)^2} = 12.60$$

③ Ordne die Instanzen den nächsten Clusterzentren zu

④ Berechne Werte für $Z1_{neu}$ und $Z2_{neu}$ mit der Formel $\vec{\mu} = \frac{1}{|C_j|} \sum_{\vec{x} \in C_j} \vec{x}$

$$Z1_{neu}^{\vec{}} = \frac{1}{2}(\vec{Strand} + \vec{Meer}) = \frac{1}{2} \begin{pmatrix} 12 \\ 7 \\ 2 \\ 0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 11 \\ 5 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 11.5 \\ 6 \\ 1 \\ 0.5 \end{pmatrix}$$

Beispiel

Distanzen	Strand	Meer	Student	Klausur	Z1	Z2
Strand	0	3.16	20.92	20.37	12.60	12.76
Meer	–	0	19.89	20.37	10.81	11.26
Student	–	–	0	7.55	15.32	11.87
Klausur	–	–	–	0	16.67	13.42

Beispiel 2: k-means Clustering

① Lege 2 zufällige Clusterzentren fest: $\vec{z}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, $\vec{z}_2 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$

② Berechne Distanzen zu Instanzen:

$$|\vec{\text{Strand}} - \vec{z}_1| = \sqrt{(12-1)^2 + (7-1)^2 + (2-1)^2 + (0-1)^2} = 12.60$$

③ Ordne die Instanzen den nächsten Clusterzentren zu

④ Berechne Werte für $Z1_{neu}$ und $Z2_{neu}$ mit der Formel $\vec{\mu} = \frac{1}{|c_j|} \sum_{\vec{x} \in c_j} \vec{x}$

$$Z1_{neu}^{\vec{z}} = \frac{1}{2}(\vec{\text{Strand}} + \vec{\text{Meer}}) = \frac{1}{2} \begin{pmatrix} 12 \\ 7 \\ 2 \\ 0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 11 \\ 5 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 11.5 \\ 6 \\ 1 \\ 0.5 \end{pmatrix}$$

Beispiel

Distanzen	Strand	Meer	Student	Klausur	Z1	Z2
Strand	0	3.16	20.92	20.37	12.60	12.76
Meer	-	0	19.89	20.37	10.81	11.26
Student	-	-	0	7.55	15.32	11.87
Klausur	-	-	-	0	16.67	13.42

Beispiel 2: k-means Clustering

① Lege 2 zufällige Clusterzentren fest: $\vec{z}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, $\vec{z}_2 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$

② Berechne Distanzen zu Instanzen:

$$|\vec{Strand} - \vec{z}_1| = \sqrt{(12-1)^2 + (7-1)^2 + (2-1)^2 + (0-1)^2} = 12.60$$

③ Ordne die Instanzen den nächsten Clusterzentren zu

④ Berechne Werte für $Z1_{neu}$ und $Z2_{neu}$ mit der Formel $\vec{\mu} = \frac{1}{|C_j|} \sum_{\vec{x} \in C_j} \vec{x}$

$$Z1_{neu}^{\vec{z}} = \frac{1}{2}(\vec{Strand} + \vec{Meer}) = \frac{1}{2} \begin{pmatrix} 12 \\ 7 \\ 2 \\ 0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 11 \\ 5 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 11.5 \\ 6 \\ 1 \\ 0.5 \end{pmatrix}$$

Beispiel

Distanzen	Strand	Meer	Student	Klausur	Z1	Z2
Strand	0	3.16	20.92	20.37	12.60	12.76
Meer	-	0	19.89	20.37	10.81	11.26
Student	-	-	0	7.55	15.32	11.87
Klausur	-	-	-	0	16.67	13.42

Beispiel 2: k-means Clustering

① Lege 2 zufällige Clusterzentren fest: $\vec{z}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, $\vec{z}_2 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$

② Berechne Distanzen zu Instanzen:

$$|\vec{\text{Strand}} - \vec{z}_1| = \sqrt{(12-1)^2 + (7-1)^2 + (2-1)^2 + (0-1)^2} = 12.60$$

③ Ordne die Instanzen den nächsten Clusterzentren zu

④ Berechne Werte für $Z1_{neu}$ und $Z2_{neu}$ mit der Formel $\vec{\mu} = \frac{1}{|C_j|} \sum_{\vec{x} \in C_j} \vec{x}$

$$Z1_{neu} = \frac{1}{2}(\vec{\text{Strand}} + \vec{\text{Meer}}) = \frac{1}{2} \begin{pmatrix} 12 \\ 7 \\ 2 \\ 0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 11 \\ 5 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 11.5 \\ 6 \\ 1 \\ 0.5 \end{pmatrix}; \quad Z2_{neu} = \begin{pmatrix} 0.5 \\ 1 \\ 12 \\ 12 \end{pmatrix}$$

Inhaltsverzeichnis

- 1 Verschiedene Arten von Clustering
- 2 Clusteringmethoden und -algorithmen
 - Hierarchisches Clustering
 - Flaches Clustering
- 3 Anwendungsbeispiel: Nach semantischer Ähnlichkeit clustern
- 4 Evaluation von Clustering Modellen
- 5 Course Wrap-up

Clusteringmodelle Evaluieren

- Teile Daten auf in Trainings- und Testdaten.
- Was bedeutet Trainingsdaten und Testdaten für unüberwachte Algorithmen?

Trainingsdaten:

Instanzen, die benutzt wurden um die Cluster zu generieren.
(z.B. Clusterzentren by k-means zu berechnen).

Testdaten:

Ungesehene Instanzen, die mit dem trainierten Clustermodell klassifiziert werden.

- **Frage:** Woher wissen wir bei flachen unüberwachten Verfahren, wie viele Cluster k wir generieren sollen?
- **Antwort:** Wir können verschiedene Werte für k ausprobieren und den Wert für k nehmen, der am besten funktioniert. → Validierungsdaten.

Woher wissen wir bei unüberwachten Verfahren eigentlich, was “richtig” ist?

- **Intuitiv** testen: sind die generierten Cluster sinnvoll? **NICHT EMPFOHLEN.**
- Einen unabhängigen **Experten** die Instanzen aus dem Testset manuell clustern lassen und mit den automatisch generierten Clustern vergleichen.
- Gegen vordefinierte **Klassifikation** testen, falls es eine solche gibt
- **Aufgaben-orientierte** Evaluation: inwiefern verbessert das Clustering die Performanz auf einer bestimmten Aufgabe?

Clusteringevaluation Beispiel

Beispiel

Wir wollen mit einem Clusteringalgorithmus automatisch POS tags lernen – der Algorithmus soll also alle Wörter die das gleiche POS tag haben in ein Cluster packen.

- **Intuitiv:** Schauen, ob die Worte, die im gleichen Cluster gelandet sind, das gleiche POS tag zu haben scheinen.
- **Experte:** einen Linguisten beauftragen, die Worte nach POS tags zu gruppieren und gegen automatische Cluster vergleichen.
- **Klassifikation:** Wörter im Wörterbuch nachschauen und sehen, ob die Wörter im gleichen Cluster im Wörterbuch auch mit gleichem POS tag vorkommen.
- **Aufgaben-orientiert:** Nutze die Clusterinformation z.B. für Parsing und teste, ob sich die Parsingperformance verbessert.

Zusammenfassung Clustering

- Clustering vs. Klassifizierung: Clustering ist unsupervised, die Klassen sind noch nicht bekannt.
- Hierarchisches vs. flaches Clustering
- Unterschiedliche Ähnlichkeitsmaße bei hierarchischem Clustering
 - single link
 - complete link
 - average link
- k-means Algorithmus für flaches Clustering
- Hier nur hartes Clustering betrachtet, desweiteren gibt es disjunktives und weiches Clustering.
- Evaluierung mit annotierten Daten oder inwiefern die Cluster bei einer anderen Aufgabe helfen.

Inhaltsverzeichnis

- 1 Verschiedene Arten von Clustering
- 2 Clusteringmethoden und -algorithmen
 - Hierarchisches Clustering
 - Flaches Clustering
- 3 Anwendungsbeispiel: Nach semantischer Ähnlichkeit clustern
- 4 Evaluation von Clustering Modellen
- 5 Course Wrap-up

Learning Bias

To generalize successfully, a machine learning system uses a *learning bias* to guide it through the space of possible concepts. Also called *inductive bias* (Mitchell, 1997).

Language bias: the language in which the result is expressed determines which concepts can be learned.

Example: a learner that can learn disjunctive rules will get different results than one that doesn't use disjunction.

An important factor is *domain knowledge*, e.g., the knowledge that certain combinations of attributes can never occur.

Learning Bias

Search bias: the way the space of possible concepts is searched determines the outcome of learning.

Example: greedy search: try to find the best rule at each stage and add it to the rule set; beam search: pursue a number of alternative rule sets in parallel.

Two common search biases are *general-to-specific* (start with a general concept description and refine) and *specific-to-general* (start with a specific example and generalize).

Learning Bias

Overfitting-avoidance bias: avoid learning a concept that overfits, i.e., just enumerates the training data: this will give very bad results on test data, as it lacks the ability to generalize to unseen instances.

Example: consider simple concepts first, then proceed to more complex one, e.g., avoid a complex rule set with one rule for each training instance.

Weiter Themen:

- Grundlagen Informationstheorie
- Anwendungen der Informationstheorie in der CL
 - Kollokationsextraktion
 - Subkategorisierungsrahmen lernen
 - Semantische Selektionspräferenzen lernen
- Maschinelles Lernen
 - Naive Bayes Klassifizierer
 - Entscheidungsbäume
 - Clustering
- Probabilistische kontextfreie Grammatiken (PCFGs)

Plan

Date	Montag 4-6pm	Date	Dienstag 4-6pm	Date	Donnerstag 2-4pm
15/06/15	No Lecture	16/06/15	Zwischenklausur	18/06/15	L2: HMMs (Ü1 austeilern)
22/06/15	L3: Entropy (Ü2 austeilern)	23/06/15	Corpora (Tut1)	25/06/15	L4: Kolloccations (Ü3 austeilern)
29/06/15	L5: PCFGs 1 (Vg + Ü4 austeilern)	30/06/15	Entropy (Tut2)	02/07/15	L6: Psycholinguistics
06/07/15	L7: Naive-Bayes	07/07/15	Kolloccations (Tut3)	09/07/15	L7: Decision Trees (Ü4 austeilern)
13/07/15	L8: Clustering (Ü5 austeilern)	14/07/15	Decision Trees (Tut4)	16/07/15	L10: PCFGs 2
20/07/15	PCFG-Ü besprechen (Tut5)	21/07/15	Probeklausur	23/07/15	Office Hour for Questions
27/07/15	Free	28/07/15	Klausur	30/07/15	Free

- Office hour: Jesus Calvillo oder Vera Demberg (Fragen erst per e-mail)

Klausur

Dienstag 28. Juli, 2015 um 16:00 s.t.

Konferenzraum, Gebäude C7.4

- 1 Es stehen 90 Minuten zur Bearbeitung zur Verfügung.
- 2 Bitte tragen Sie Ihren Namen und Ihre Matrikelnummer in das obige Feld ein.
- 3 Erlaubte Hilfsmittel sind **ein** Blatt selbsterstellte (handgeschriebene) Zusammenfassung des Stoffes und einen Taschenrechner (kein Handy!).
- 4 Bitte keinen Bleistift benutzen!
- 5 Die einzelnen Lösungsschritte bitte soweit begründen, dass der Lösungsansatz nachvollziehbar ist.
- 6 Insgesamt können 50 Punkte erreicht werden.