Karlsruher Institut für Technologie

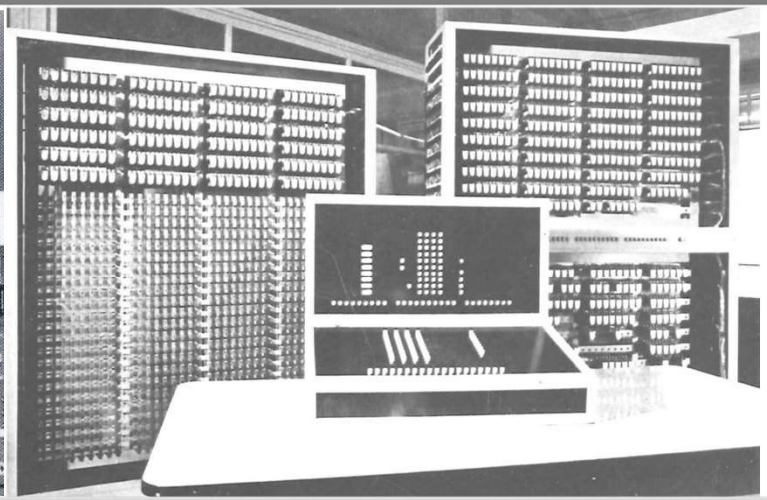
Institut für Technik der Informationsverarbeitung

Institutsleitung
Prof. Dr.-Ing. J. Becker
Prof. Dr.-Ing. E. Sax
Prof. Dr. rer. nat. W. Stork

6.2 Big Data & Maschinelles Lernen (C)

Informationstechnik II

Prof. Dr.-Ing. Eric Sax

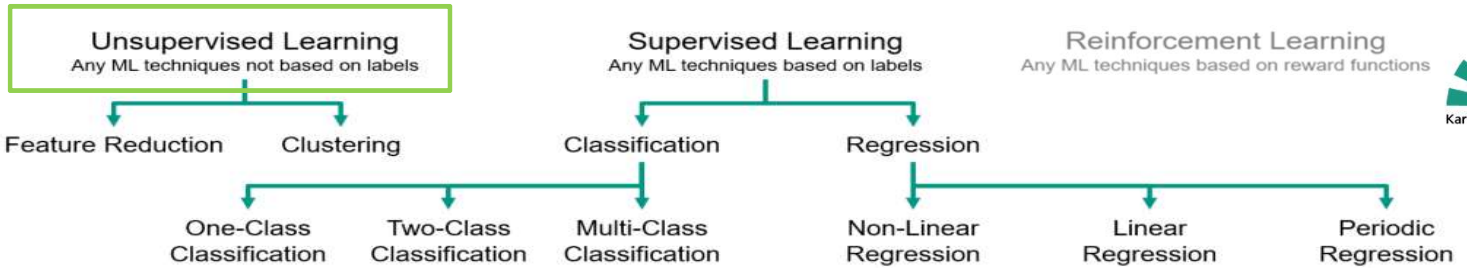


Inhalt IT2

6. Big Data & Maschinelles Lernen

- Definitionen, Verwendung (Charakteristik, Risiken, Chancen)
- Motivation und Anwendungsfälle
- Data Science Prozesse:
  - KDD
  - CRISP-DM
- CRISP-DM im Detail
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
- Infrastruktur für Big Data

Machinelles Lernen

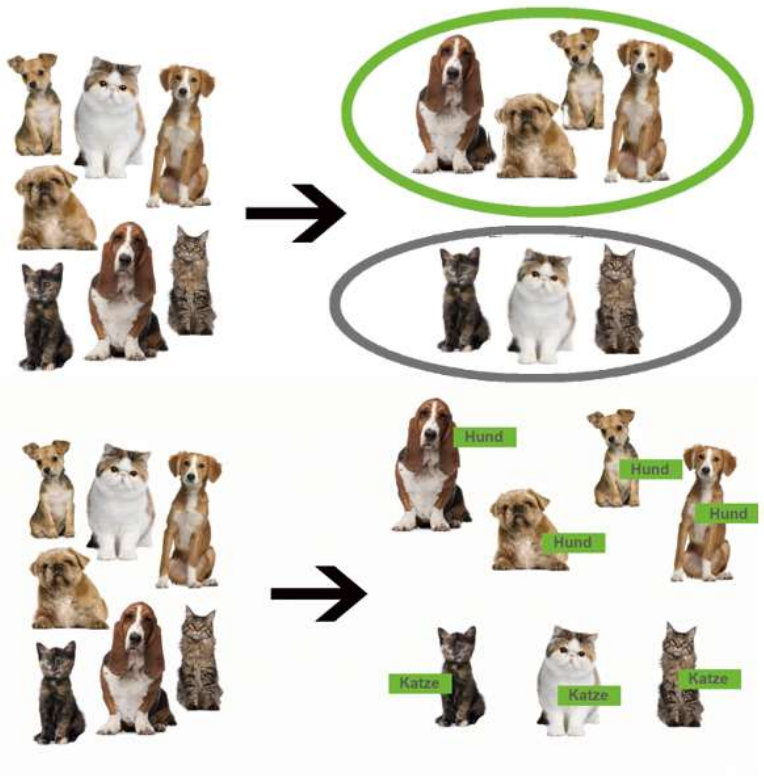


Algorithm ↓	Abk.	Feature Reduction	Clustering	One-Class Classification	Two-Class Classification	Multi-Class Classification	Non-Linear Regression	Linear Regression	Periodic Regression
Faktorenanalyse		X							
Principal Component Analysis	PCA	X		X					
	K-means		X						
hierarchical cluster analysis	HCA		X						
DBSCAN			X						
One Class Support Vector Machine	OCSVM			X					
Isolation Forest				X					
	LODA			X					
(künstliche) Neuronale Netze	NN			X (Autoencoder)	X	X	X	X	X
Support Vector Machine	SVM				X				
Decision Tree					X	X			
Bayes-Klassifikation					X	X			
Random Forest						X			
Diskriminanzanalyse				x	x	x			
Logistic Regression							X	X	
Linear Regression								X	
Harmonic Regression									X
Nächste-Nachbar-Klassifikation	K-NN	x	x						

Überwachtes vs. Unüberwachtes Lernen

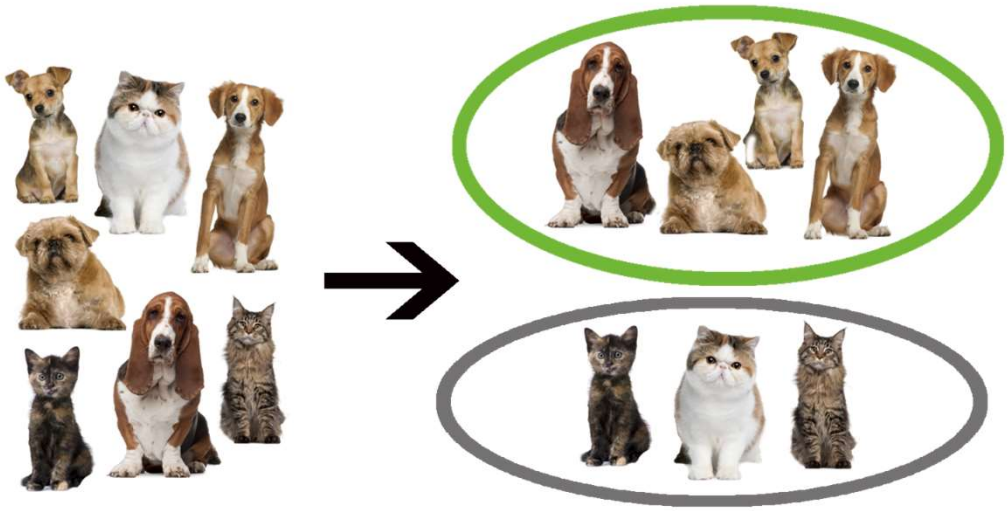
Beispiel: Hunde und Katzen

- Unüberwachtes Lernen:
  - Das System weiß nicht, was es erkennen soll.
  - Wenn es Bilder von Tieren verarbeitet, teilt das System alles was aussieht wie eine Katze oder alles was aussieht wie ein Hund in entsprechende Gruppen ein, ohne diese jedoch so zu benennen, da nicht definiert ist, was eine Katze und was ein Hund ist.
  - Diese Methode wird angewandt, wenn Daten noch unbekannt sind und entsprechend keine Vorgaben vorhanden sind.
- Überwachtes Lernen:
  - Es gibt Trainings-Daten, bei denen wir die Eingangs-Parameter sowie das Ergebnis kennen.
  - Aus den Trainings-Daten werden Modelle von Hunden und Katzen erstellt, die zusammen mit den Machine Learning Algorithmen das Ergebnis liefern.
  - Da die Modelle erstellt wurden, kann man „unbekannte“ Bilder liefern und das System berechnet dafür das Ergebnis (Prediction).



2 Ansätze des Unüberwachten Lernen

- Clustering
  - Der Machine Learning Algorithmus wird die Bilder in die Kategorien Katzen und Hunde aufteilen, obwohl er darauf nicht trainiert wurde und ohne die Kategorien zu kennen.

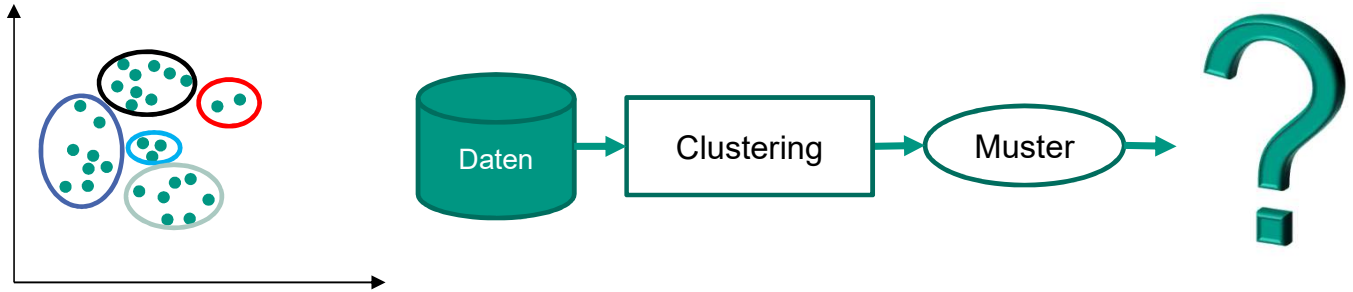


- Reduzierung der Dimensionalität (dimensionality oder feature reduction)
  - Das System erkennt durch ein Muster, welche (im Vorfeld definierten) Dimensionen zusammen gehören und reduziert diese.
  - Das steigert die Performance bei der Verarbeitung und Analyse.

Modellierung der Daten – Data Mining

Clustering (unüberwacht)

- Gruppierung von gleichartigen Situationen

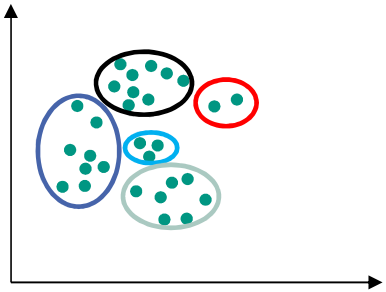


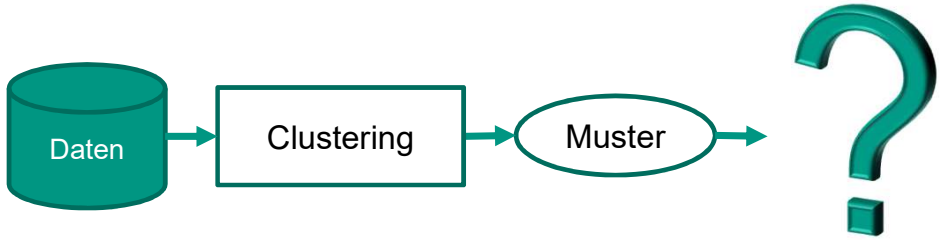
- Vorteil:
  - Benötigt keine vorherigen Annotationen
  - Entdeckung von bisher unbekannten Differenzierbarkeiten
- Nachteil
  - Nicht zielgerichtet
  - Stark von Datenselektion, Transformation und Clustering-Verfahren abhängig
  - Bewertung der Ergebnisse schwierig

Modellierung der Daten – Data Mining

Clustering (unüberwacht)

- Gruppierung von gleichartigen Situationen

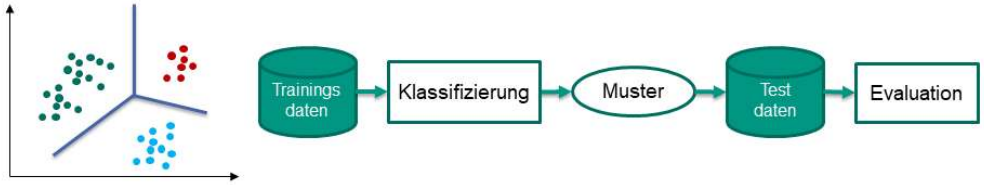




- Vorteil:
  - Benötigt keine vorherigen Anr
  - Entdeckung von bisher unbek
- Nachteil
  - Nicht zielgerichtet
  - Stark von Datenselektion, Tra
  - Bewertung der Ergebnisse sc

Support Vector Machine (überwacht)

- Gruppierung/ Separierung von vorsortierten Daten



- Vorteil:
  - Zielgerichtet durch Vorsortierung
  - Evaluation durch Testdaten möglich
    - Vergleich der entdeckten Muster (Optimierung)
    - Vergleich erleichtert Merkmalselektion
- Nachteil:
  - Nur anwendbar auf vorsortierte Daten
  - Neues Wissen beschränkt auf die vorgegeben Klassen

Clusteranalyse

Unüberwachtes Lernen, ohne Belohnung, ohne Vorwissen

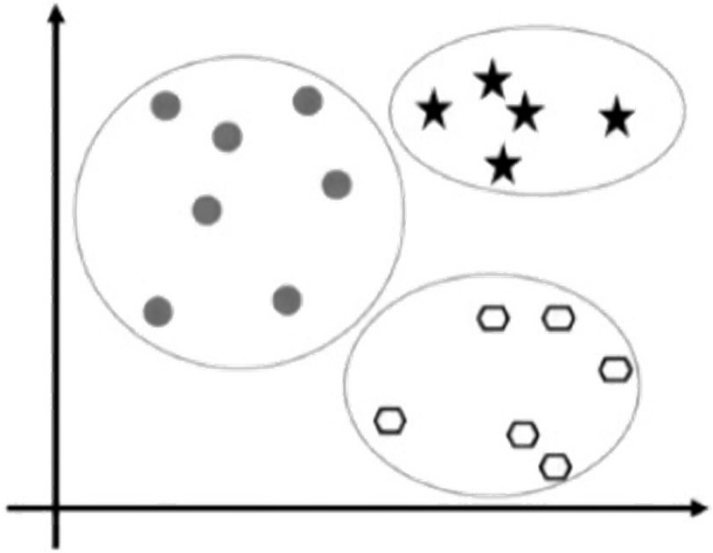
- Clustering nimmt eine Gruppierung anhand von Ähnlichkeiten vor
  - *Dabei ist jedoch nicht gewährleistet, dass diese Ähnlichkeiten aussagekräftig oder für irgendeinen Zweck nützlich sind.*
- Ziel ist die Objekte so zu Gruppen (*Clustern*) zusammenzufassen, dass die Objekte in einer Gruppe ähnlich und die Gruppen untereinander unähnlich sind.
- Da im Gegensatz zu klassifizierenden Verfahren die Gruppen vor der Anwendung der Analyse nicht bekannt sind, handelt es sich um ein Verfahren des unüberwachten Lernens.
- Clusteranalysen werden häufig eingesetzt als:
  - Grundlage einer Markt- bzw. Kundensegmentierung,
  - Grundlage zur anschließenden, automatischen Klassifizierung von Daten,
  - Im Bereich der Bilderkennung.

Clusteranalyse

Unüberwachtes Lernen, ohne Belohnung, ohne Vorwissen

- Clustering nimmt eine Gruppierung anhand von Ähnlichkeiten vor
  - *Dabei ist jedoch nicht gewährleistet, dass diese Ähnlichkeiten aussagekräftig oder für irgendeinen Zweck nützlich sind.*
- Ziel ist die Objekte so zu Gruppen (*Clustern*) zusammenzufassen, dass die Objekte in einer Gruppe ähnlich und die Gruppen untereinander unähnlich sind.
- Da im Gegensatz zu klassifizierenden Verfahren die Gruppen vor der Anwendung der Analyse nicht bekannt sind, handelt es sich um ein Verfahren des unüberwachten Lernens.
- Clusteranalysen werden häufig eingesetzt als:
  - Grundlage einer Markt- bzw. Kundensegmentierung,
  - Grundlage zur anschließenden, automatischen Klassifizierung von Daten,
  - Im Bereich der Bilderkennung.

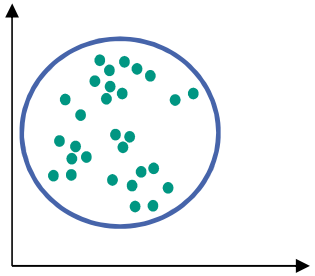
- Punktewolken
  - Die Punkte stellen die einzelnen Datensätzen in einem n-dimensionalen Raum dar, wobei die n-Dimensionen den Variablen entsprechen
  - Ein Cluster stellt dann eine Punktewolke dar, die ähnliche Punkte zusammenfasst.



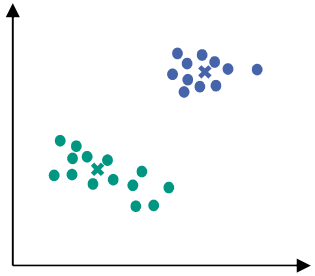
Clusteranalyse
Kategorisierung

- Unter dem Begriff der Clusteranalyse wird ein Vielzahl von verschiedenen heuristischen Verfahren zusammengefasst.
- Die einzelnen Clusterverfahren unterscheiden sich anhand der verwendeten Ähnlichkeitsmaße sowie der Vorgehensweise, mit der eine möglichst gute, eindeutige Trennung der Cluster erzielt werden soll.
  - ► Es gibt keinen Algorithmus, der eine "optimale" Gruppierung garantiert
- Die Clusterverfahren werden wie folgt unterteilt:

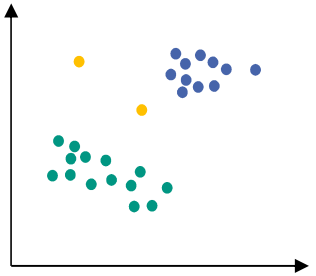
Hierarchisch



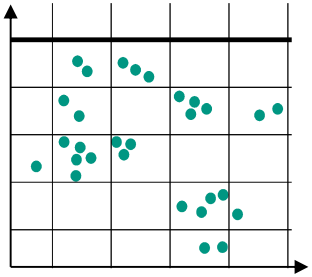
Partitionierend



Dichte-basiert



Gitter-basiert



Hierarchische Clusterverfahren

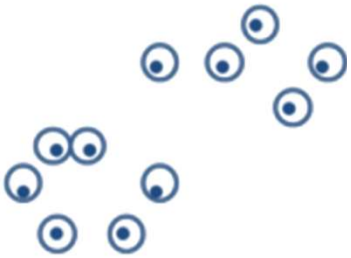
- Hierarchische Verfahren erzeugen eine Clusterstruktur möglicher Einteilungen, so dass die Objekte auf oberster Ebene in nur wenige Cluster unterteilt werden, die sich wiederum in eigene Cluster unterteilen etc.
  - Bei den agglomerierenden (anhäufenden) Verfahren wird initial jedes Objekt der Datensammlung als eigenes Cluster interpretiert.
    - Es werden sukzessive einzelne, zueinander ähnliche Objekte zu Clustern und bereits erkannte Cluster zu größeren Clustern verschmolzen bis der Abstand zwischen je zwei Clustern einen bestimmten Schwellwert übersteigt oder wenn nur noch ein großes Cluster übrig ist.
  - Divisive (unterteilende) Verfahren beginnen anfänglich mit einem einzigen Cluster.
    - Jedes Objekt der Datensammlung ist diesem Cluster zugeordnet. Anschließend werden immer neue (Teil-) Cluster erstellt, d.h. das ursprüngliche Cluster immer feiner unterteilt.
    - Dieser Prozess dauert so lange an, bis nur noch Cluster bestehend aus Einzelobjekten übrig bleiben.
- Hierarchische Clusterverfahren sind relativ rechenaufwändig.
  - Sie besitzen eine Laufzeit von etwa $O(n^2)$.
  - Durch Verwendung zusätzlicher Heuristiken kann jedoch eine annähernd lineare Laufzeit erreicht werden.

Quelle: GRIN - Clustering und Evaluierung von Benutzerprofilen bei Web-Portalen https://www.grin.com/document/56106

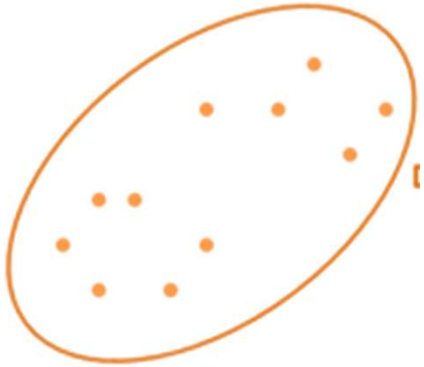
Hierarchische Clusterverfahren

Agglomerativ vs. *Divisiv*

Agglomerative Hierarchical Clustering

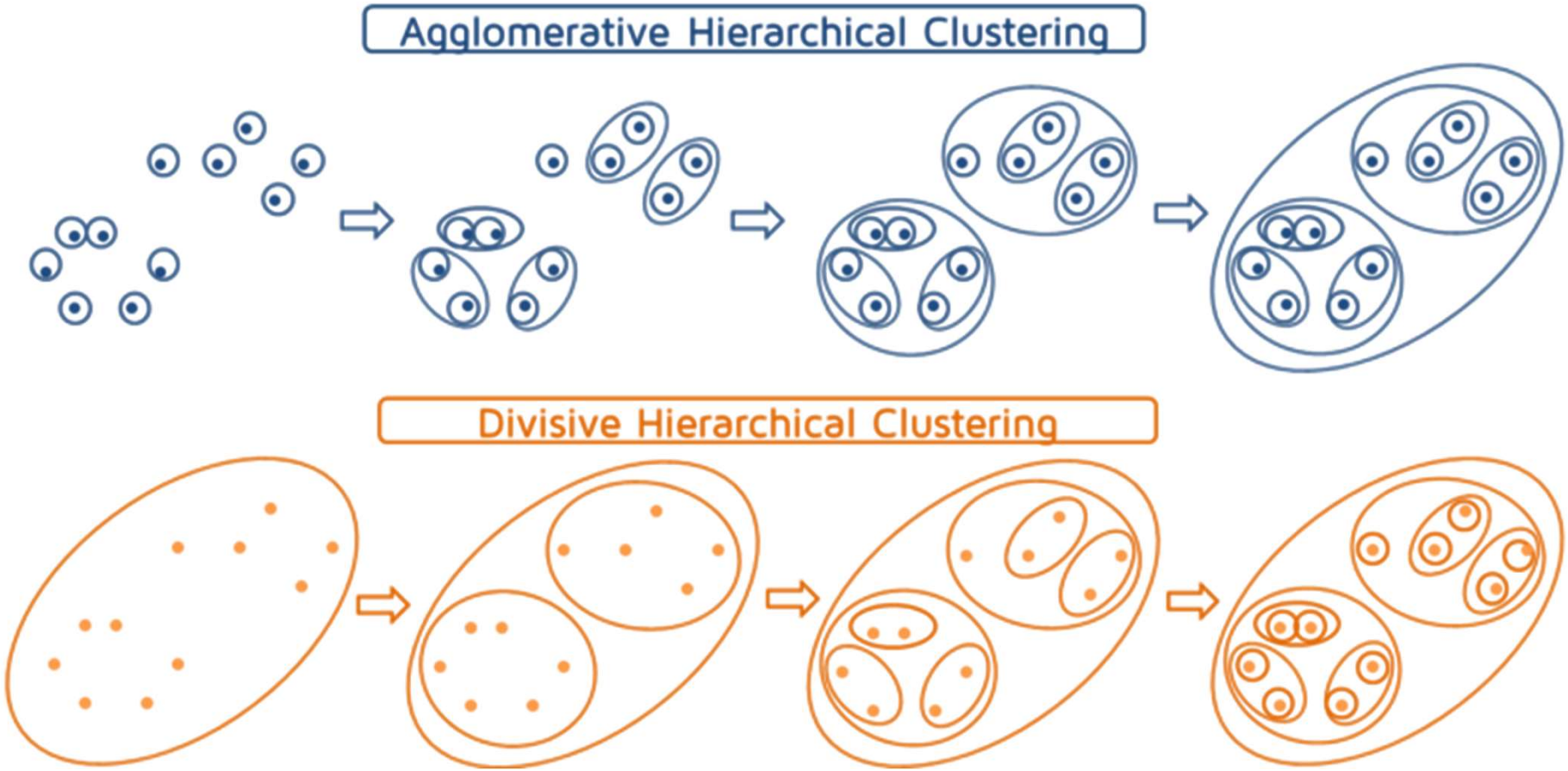


Divisive Hierarchical Clustering



Hierarchische Clusterverfahren

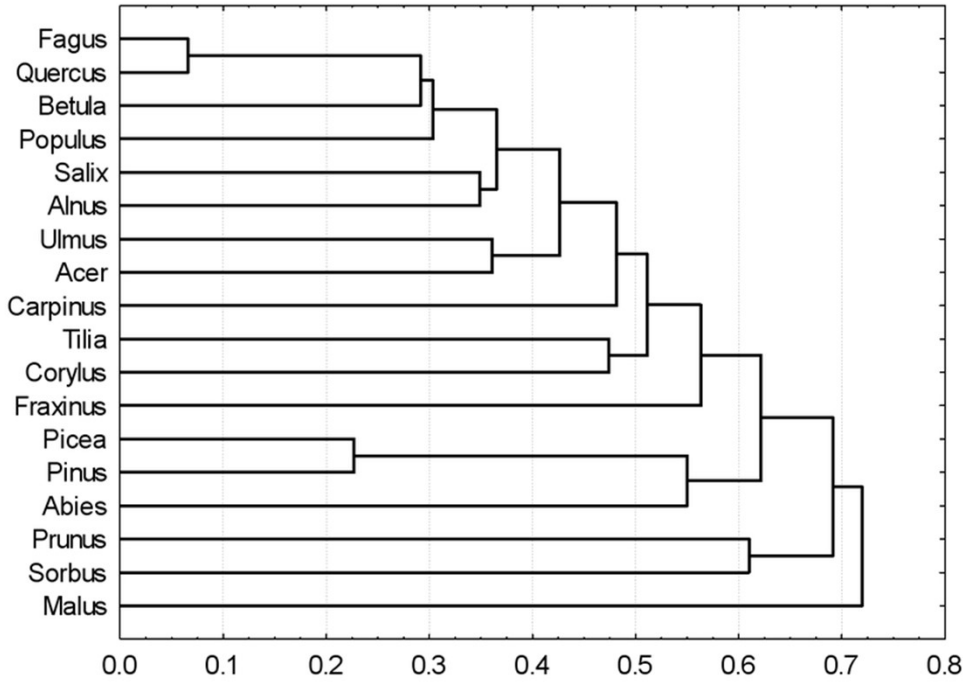
Agglomerativ vs. *Divisiv*



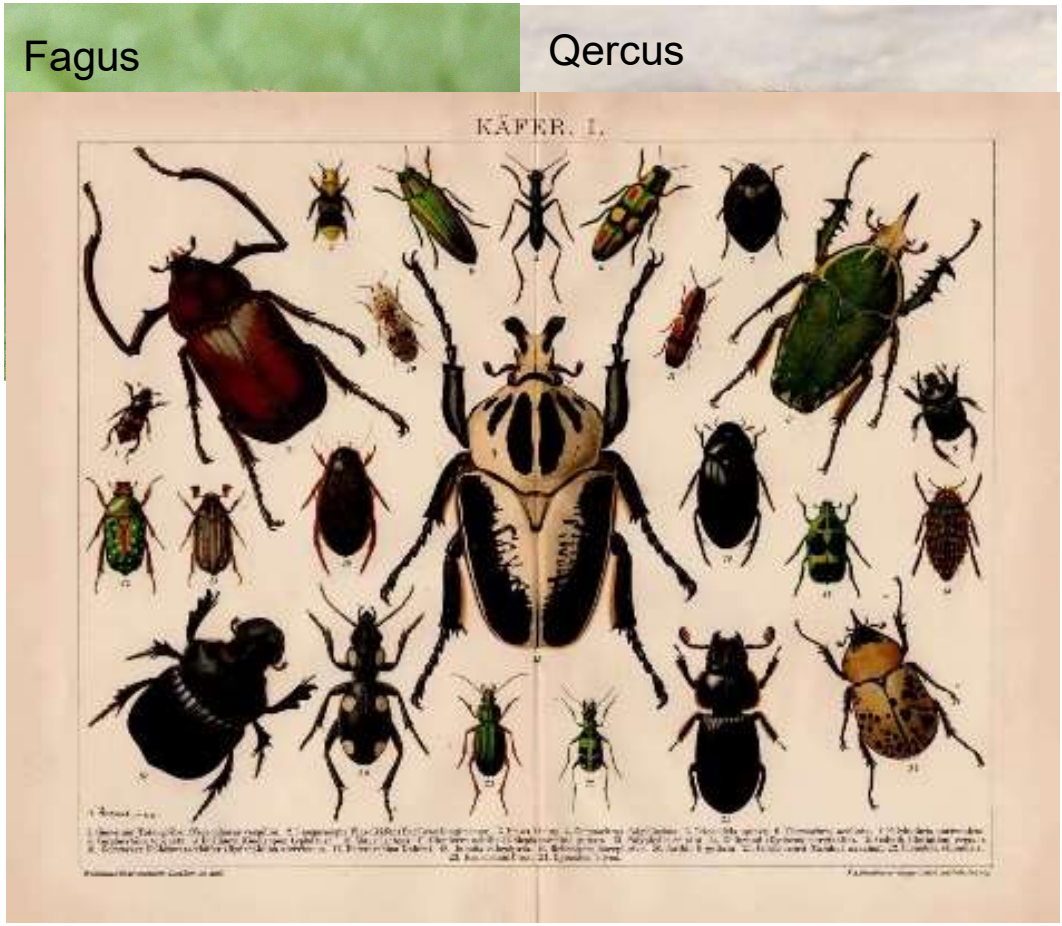
Dendrogramm

- Beiden hierarchischen Verfahrenstypen, *agglomerierend* wie *divisiv*, ist gemeinsam, dass eine Baumstruktur entsteht, die meist als Dendrogramm bezeichnet wird.
- Die Länge eines Querbalkens zeigt an, auf welcher Hierarchieebene Cluster verschmolzen werden.
- Wenn ein einzelner Punkt erst weit oben im Dendrogramm mit den anderen Gruppierungen vereint wird, weist dies darauf hin, dass er sich deutlich von den anderen Punkten unterscheidet.
  - Wir bezeichnen das als einen »Ausreißer«

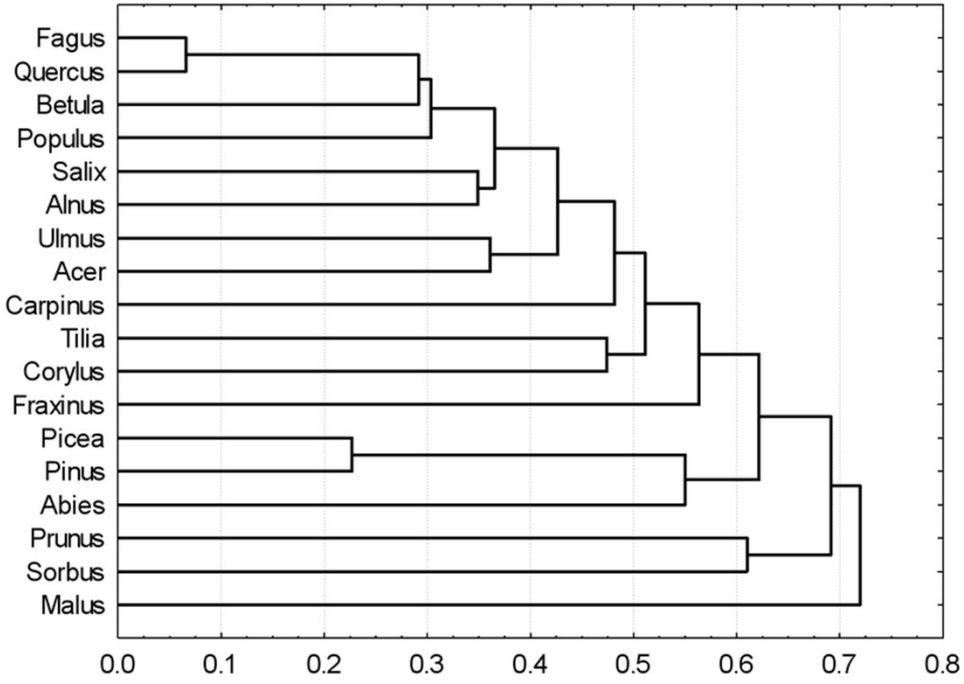
- Beispiel: *Dendrogramm der in einer Clusteranalyse errechneten Ähnlichkeitswerte (sörensen-koeffizient) von 300 xylobionten Käferarten mit Vorkommen an Rotbuche aus dem Steigerwald und dem Spessart*



Dendrogramm

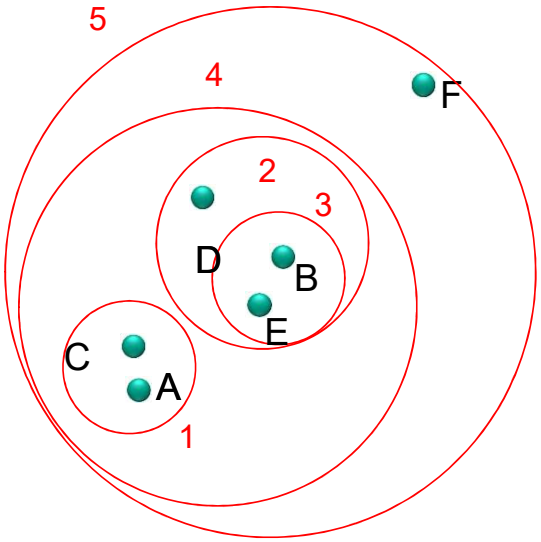


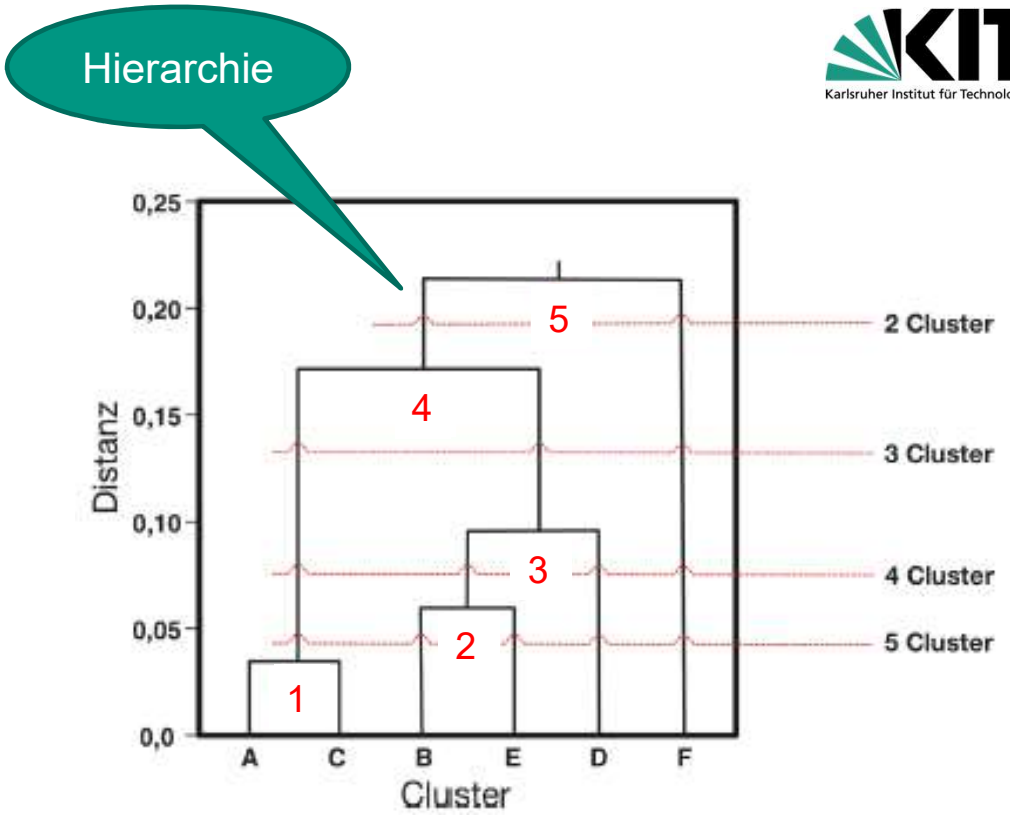
- Beispiel: *Dendrogramm der in einer Clusteranalyse errechneten Ähnlichkeitswerte (sörensen-koeffizient) von 300 xylobionten Käferarten mit Vorkommen an Rotbuche aus dem Steigerwald und dem Spessart*



Dendrogramm

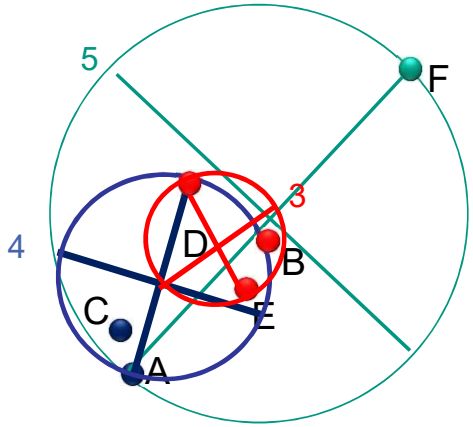
Beispiel (agglomerierend)

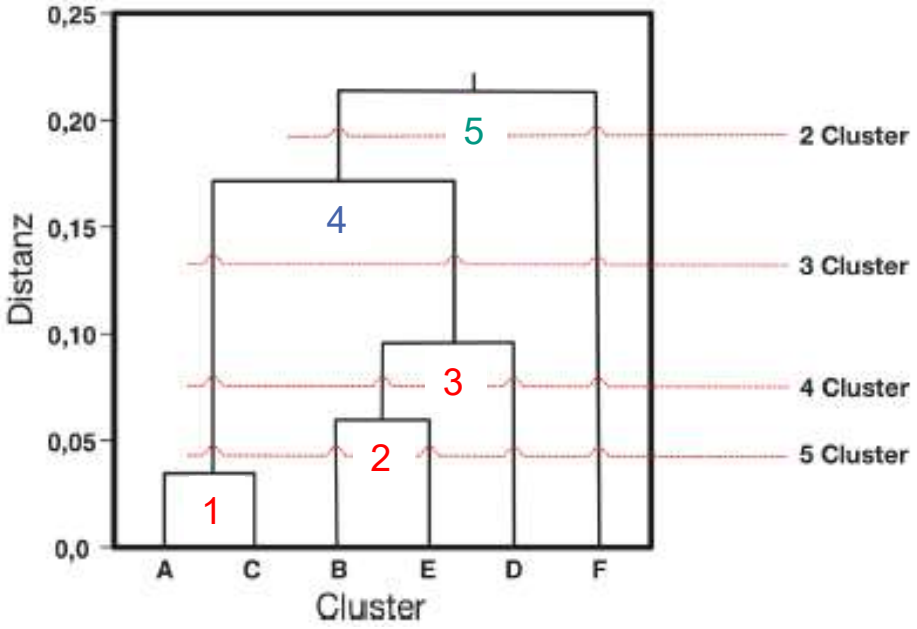




- Sechs Punkte und ihre möglichen Cluster.
- Links sind die sechs Punkte A bis F und die Kreise 1 bis 5 dargestellt, die verschiedene auf der Distanz beruhende Gruppierungen kennzeichnen.
- Sie bilden eine implizite Hierarchie → Das Dendrogramm der Gruppierung im unteren Teil der Abbildung zeigt diese Hierarchie explizit.

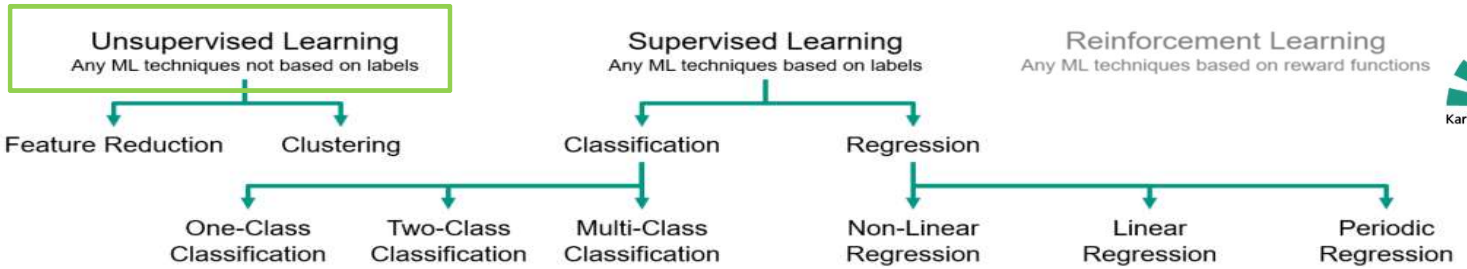
Dendrogramm
Beispiel (divisiv)





- Kaufman und Rousseeuw (1990) beschreiben eine *Divisive Clustering Procedure* wie folgt:
  - Starte mit einem kreisförmigen Cluster, das alle Punkte enthält.
    - Durchmesser ist der größte Abstand 2er Punkte (Maximaldistanz)
  - Dieser Cluster wird in zwei Cluster geteilt, wobei die Senkrechte auf die Mitte des Abstandes den Teiler bildet.
  - Die Schritte werden je Teil-Cluster solange wiederholt, bis alle Cluster nur noch ein Objekt enthalten.

Machinelles Lernen



Algorithm	Abk.	Feature Reduction	Clustering	One-Class Classification	Two-Class Classification	Multi-Class Classification	Non-Linear Regression	Linear Regression	Periodic Regression
Faktorenanalyse		X							
Principal Component Analysis	PCA	X		X					
	K-means		X						
hierarchical cluster analysis	HCA		X						
DBSCAN			X						
One Class Support Vector Machine	OCSVM			X					
Isolation Forest				X					
	LODA			X					
(künstliche) Neuronale Netze	NN			X (Autoencoder)	X	X	X	X	X
Support Vector Machine	SVM				X				
Decision Tree					X	X			
Bayes-Klassifikation					X	X			
Random Forest						X			
Diskriminanzanalyse				x	x	x			
Logistic Regression							X	X	
Linear Regression								X	
Harmonic Regression									X
Nächste-Nachbar-Klassifikation	K-NN	x	x						

Partitionierende Clusterverfahren

- Die Aufgabe partitionierender Clusterverfahren besteht darin, eine Datensammlung, ausgehend von einer initialen Partitionierung, in k disjunkte Mengen derart zu partitionieren, dass sich die Objekte innerhalb einer Gruppe so ähnlich wie möglich sind.
  - Jedes Objekt wird einem eindeutigen Cluster zugewiesen.
  - Es entsteht keine hierarchische Clusterstruktur.
- Der Vorteil partitionierender Clusterverfahren liegt in der Untersuchung sehr großer Datensammlungen, wo die Erstellung eines Dendrogramms nur schwer durchzuführen ist.
- *Bei partitionierenden Clusterverfahren ist es notwendig, aber auch problematisch, vor dem Start des Algorithmus anzugeben, auf wie viele (unbekannte) Partitionen k der Algorithmus die Datensammlung untersuchen soll.*
- Damit bleibt die Anzahl der Cluster konstant.
  - Sicherlich lässt sich der Algorithmus mehrere Male mit verschiedenen Werten für k starten, jedoch muss man in der Lage sein, sich zwischen verschiedenen k-Werten zu entscheiden.
  - Welches k zur optimalen Clustereinteilung führt, kann nur anhand einer Ähnlichkeitsfunktion (*score function*) bestimmt werden.

Partitionierende Clusterverfahren

- Anwendung
  - Partitionierende Clusterverfahren sind geeignet, um für eine vorgegebene Anzahl k von Clustern die beste Aufteilung der Objekte zu finden.
  - Deshalb ist es wichtig, eine gute Anfangsgruppierung vorzugeben.
  - Hierfür wird häufig die Lösung eines hierarchischen Clusterverfahrens verwendet, die dann mittels eines partitionierenden Verfahren verbessert wird.
  - Sinnvoll ist auch, die anfängliche Gruppierung zu variieren und die erzielten Ergebnisse zu vergleichen.
  - Auch kann die Anzahl der Cluster verändert werden
- Das wohl wichtigste partitionierende Verfahren ist der k-means Algorithmus von MacQueen.
  - Dieser wird häufig zur Verbesserung der Lösung eines hierarchischen Clusterverfahrens eingesetzt.

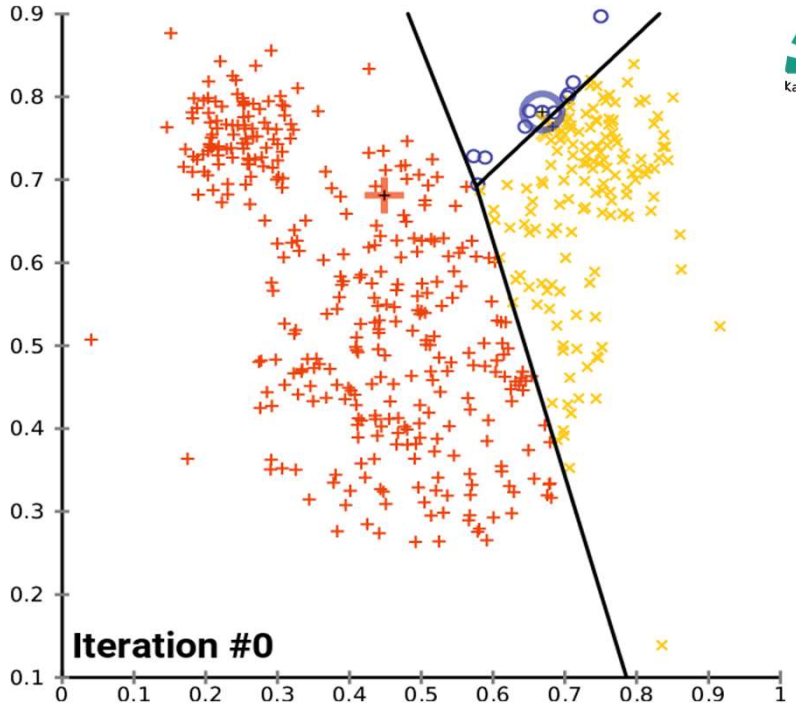
k-Means

Partitionierend

- Bildung von Mengen ähnlicher Objekte in k Gruppen
  - mit möglichst geringer Varianz
  - In Gruppen ähnlicher Größen
- Die Clusterzentren werden zufällig festgelegt und die Summe der quadrierten Abstände der Objekte zu ihrem nächsten Clusterzentrum wird minimiert.
- Das Update der Clusterzentren geschieht durch Mittelwertbildung aller Objekte in einem Cluster.

$$J = \sum_{i=1}^{k} \sum_{x_j \epsilon S_i} \left\| x_j - \mu_i \right\|^2$$

x_j: Datenpunkt
S_i: Cluster
μ_i: Schwerpunkt



Lloyd Algorithmus:

1. k unterschiedliche Zentren $c_1, c_2, \ldots, c_k$
2. Solange sich die Zielfunktion verbessert:
   Partitioniere P in Cluster $S_1, S_2, \ldots, S_k$ dass S_i die Punkte aus P enthält, deren nächstgelegenes Zentrum c_i ist
   Für jedes $1 \leq i \leq k$ sei $c_i \leftarrow \mu(C_i)$

k-Means

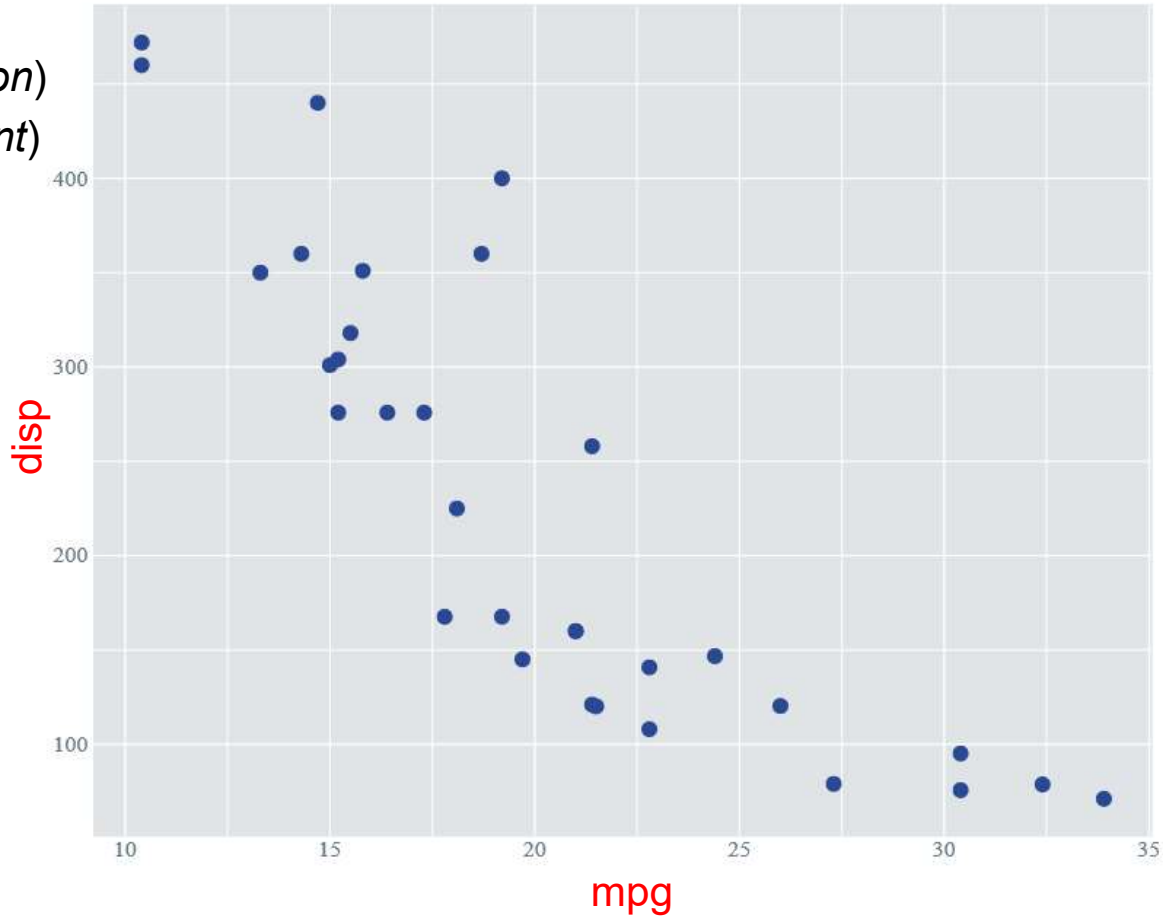
Varianten <zur Vertiefung für Interessierte>

- k-Means++
  - Als Clusterzentren werden auch zufällig Objekte so ausgewählt, so dass sie etwa uniform im Raum der Objekte verteilt sind. Dies führt zu einem schnelleren Algorithmus.
- k-Median-Algorithmus
  - Hier wird die Summe der Manhattan-Distanzen der Objekte zu ihrem nächsten Clusterzentrum minimiert. Das Update der Clusterzentren geschieht durch die Berechnung des Medians aller Objekte in einem Cluster. Ausreißer in den Daten haben dadurch weniger Einfluss.
- k-Medoids oder Partitioning Around Medoids (PAM)
  - Die Clusterzentren sind hier immer Objekte. Durch Verschiebung von Clusterzentren auf ein benachbartes Objekt wird die Summe der Distanzen zum nächstgelegenen Clusterzentrum minimiert. Im Gegensatz zum k-Means Verfahren werden nur die Distanzen zwischen den Objekten benötigt und nicht die Koordinaten der Objekte.
- EM-Clustering
  - Die Cluster werden als multivariate Normalverteilungen modelliert.
  - Mit Hilfe des EM-Algorithmus werden die unbekannten Parameter der Normalverteilungen iterativ geschätzt.
  - Im Gegensatz zu k-means wird damit eine weiche Clusterzuordnung erreicht:
  - Mit einer gewissen Wahrscheinlichkeit gehört jedes Objekt zu jedem Cluster und jedes Objekt beeinflusst so die Parameter jeden Clusters.
- Fuzzy C-Means
  - Für jedes Objekt wird ein Zugehörigkeitsgrad zu einem Cluster berechnet, oft aus dem reellwertigen Intervall [0,1] (Zugehörigkeitsgrad=1: Objekt gehört vollständig zu einem Cluster, Zugehörigkeitsgrad=0: Objekt gehört nicht zu dem Cluster).
  - Dabei gilt: je weiter ein Objekt vom Clusterzentrum entfernt ist, desto kleiner ist auch sein Zugehörigkeitsgrad zu diesem Cluster.
  - Wie im k-Median-Verfahren werden die Clusterzentren dann verschoben, jedoch haben weit entfernte Objekte (kleiner Zugehörigkeitsgrad) einen geringen Einfluss auf die Verschiebung als nahe Objekte.
  - Damit wird auch eine weiche Clusterzuordnung erreicht: Jedes Objekt gehört zu jedem Cluster mit einem entsprechenden Zugehörigkeitsgrad.

Beispiel: 32 Automobile aus dem Datensatz *mtcars*

Beispiel: s. https://www.inwt-statistics.de/blog-artikel-lesen/Clusteranalyse_in_R.html

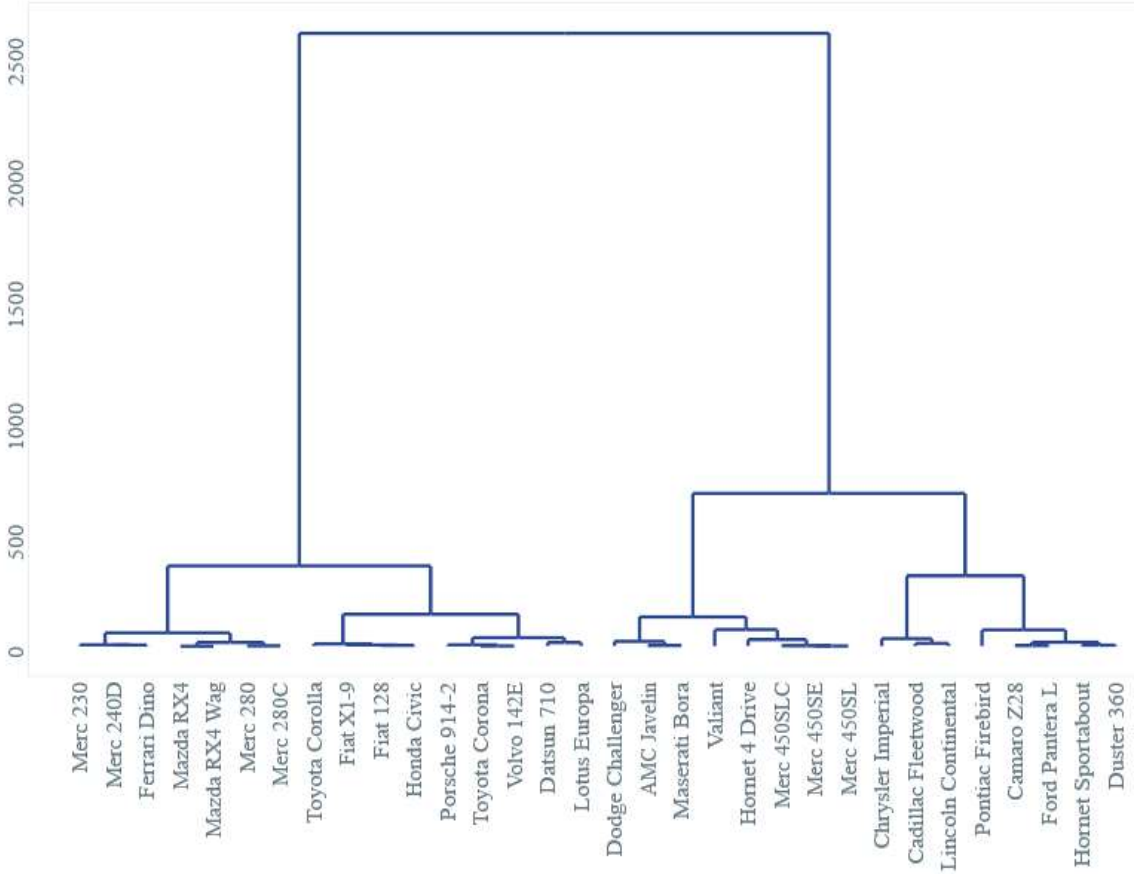
- Für die Analyse verwendete Variablen
  - mpg (Kraftstoffverbrauch, *miles per gallon*)
  - disp (Hubraum in Kubikinch, *displacement*)



Dendrogramm zu *mtcars*

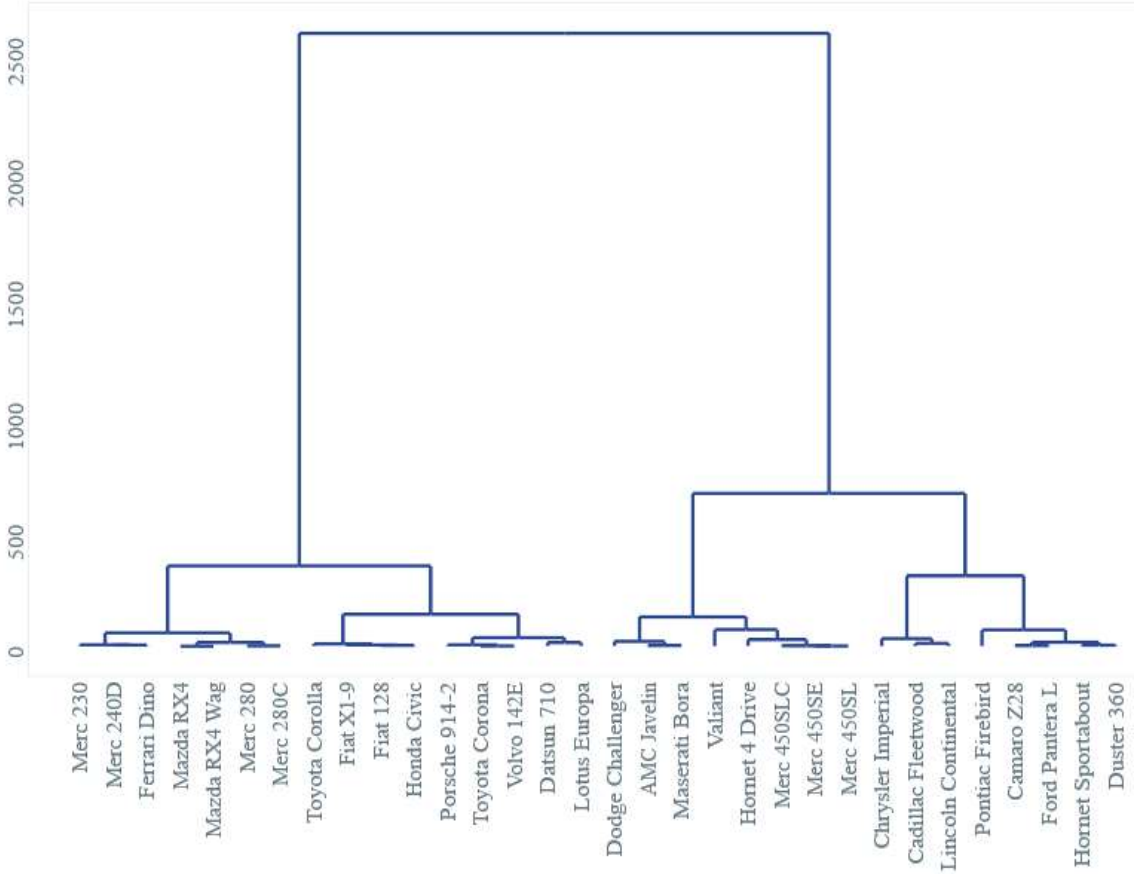
Hierarchisches Verfahren

- Für die Analyse verwendete Variablen
  - mpg (Kraftstoffverbrauch in *miles per gallon*)
  - disp (Hubraum in *Kubikinch*)
- Nun führen wir ein hierarchisches Verfahren durch
  - Auf der y-Achse des Dendrogramms ist die Distanz (euklidischer Abstand) abgetragen.
- Augenscheinlich sind 2 Cluster in diesem Falle sinnvoll.



Dendrogramm zu *mtcars*

Ferrari	
 *Ferrari Dino 246 GT*	
Dino 206 GT / 246 GT	
Produktionszeitraum:	1969–1974
Klasse:	Sportwagen
Karosserieversionen:	Coupé, Cabriolet
Motoren:	Ottomotoren: 2,0–2,4 Liter (132–143 kW)
Länge:	4200–4210 mm
Breite:	1700 mm
Höhe:	1115 mm
Radstand:	2280–2340 mm
Leergewicht:	1000–1100 kg

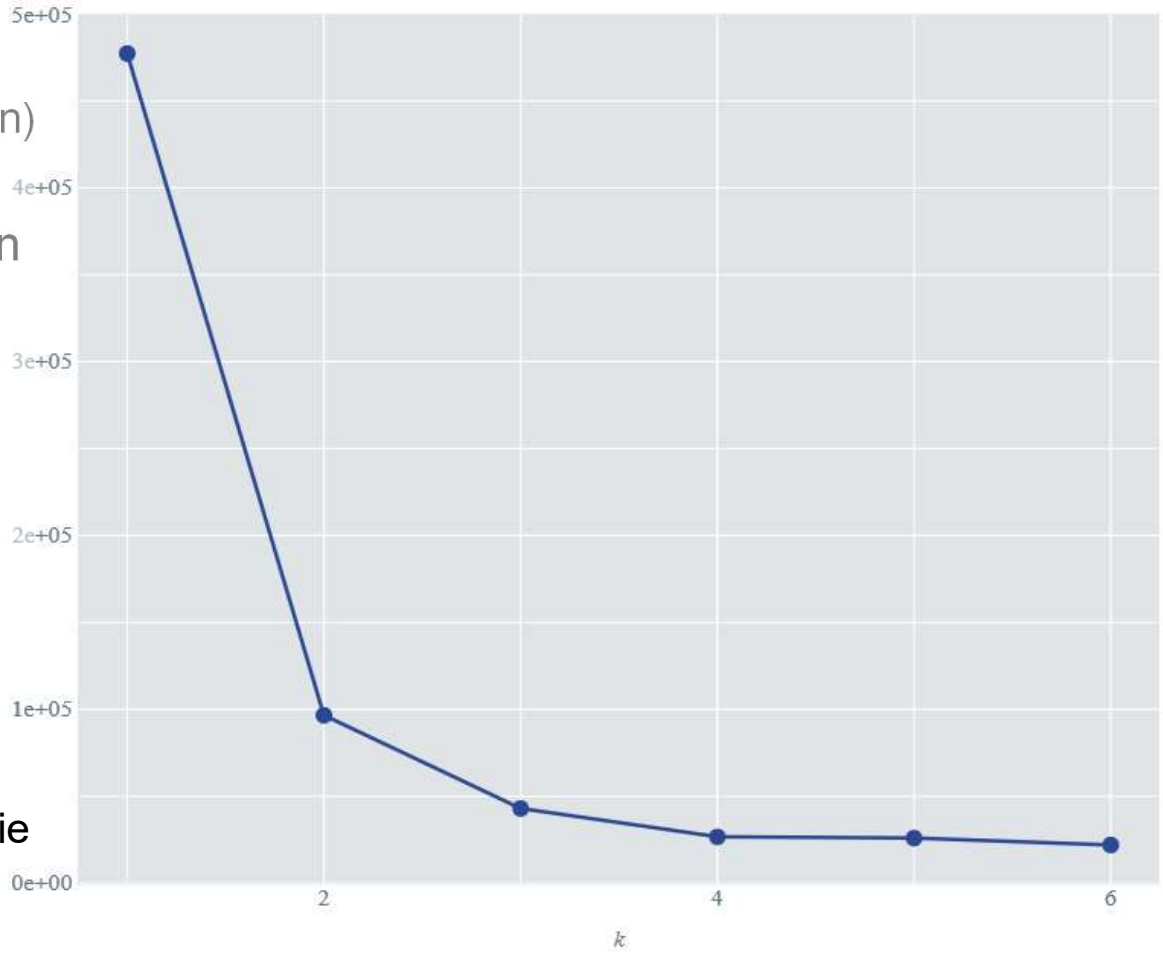


John Barry (Filmkomponist)

K-Means

Partitionierendes Verfahren

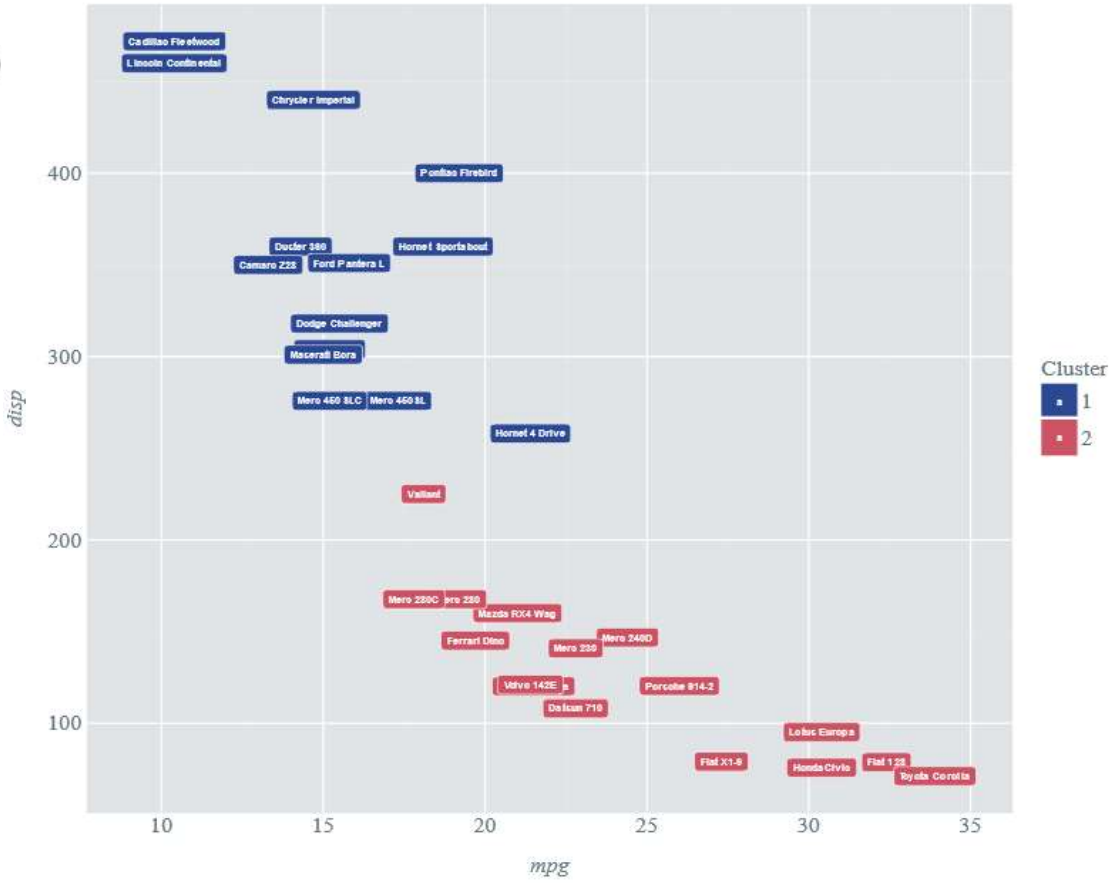
- Für die Analyse verwendete Variablen
  - mpg (Kraftstoffverbrauch in miles per gallon)
  - disp (Hubraum in Kubikinch)
- Nun führen wir ein hierarchisches Verfahren durch
  - Auf der y-Achse des Dendrogramms ist die Distanz (euklidischer Abstand) abgetragen.
- Augenscheinlich sind 2 Cluster in diesem Falle sinnvoll.
- Als nächstes erstellen wir einen Screeplot, wobei für jedes k ein k-means Verfahren durchgeführt wird.
  - Der Screeplot zeigt auf der x-Achse die Anzahl der Cluster k und auf der y-Achse die Varianz innerhalb der Cluster.



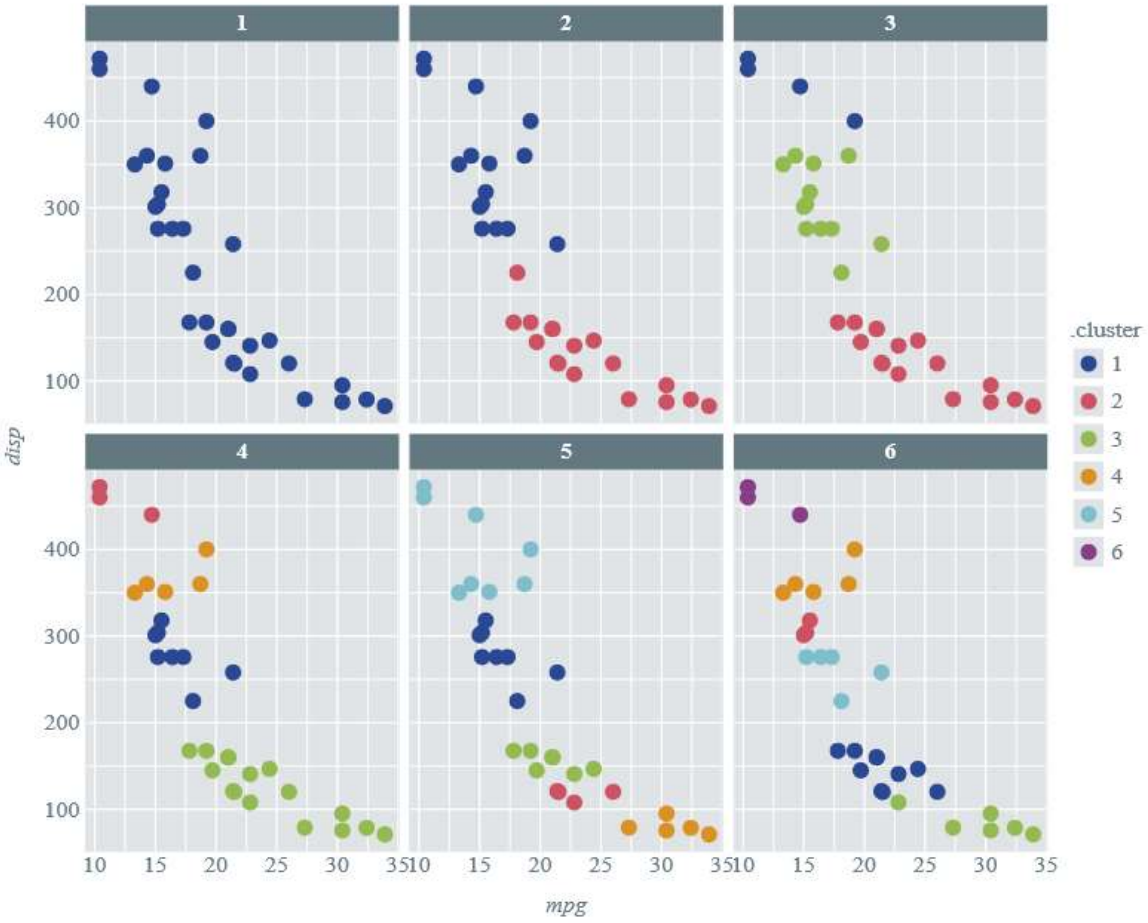
K-Means

Partitionierendes Verfahren

- Für die Analyse verwendete Variablen
  - mpg (Kraftstoffverbrauch in *miles per gallon*)
  - disp (Hubraum in *Kubikinch*)
- Nun führen wir ein hierarchisches Verfahren durch
  - Auf der y-Achse des Dendrogramms ist die Distanz (euklidischer Abstand) abgetragen.
- Augenscheinlich sind 2 Cluster in diesem Falle sinnvoll.
- Nun sehen wir uns die durch den k-means-Algorithmus erhaltene Gruppenstruktur für k=2 an



- Für die Analyse verwendete Variablen
  - mpg (Kraftstoffverbrauch in *miles per gallor*
  - disp (Hubraum in *Kubikinch*)
- Nun führen wir ein hierarchisches Verfahren durch
  - Auf der y-Achse des Dendrogramms ist die Distanz (euklidischer Abstand) abgetragen.
- Augenscheinlich sind 2 Cluster in diesem Falle sinnvoll.
- Nun sehen wir uns die durch den k-means-Algorithmus erhaltene Gruppenstruktur für k=2 an
- Vergleichsweise die Gruppenstruktur auch für k=1, k=3 etc.



Interpretation der Clusterlösung

- Nun zum eigentlich schwierigsten Teil der Analyse: der Interpretation.
  - Für die Analyse haben wir den Kraftstoffverbrauch mpg und den Hubraum disp herangezogen.
- Zur Hilfestellung nehmen wir die restlichen Informationen (s. rechts oben) aus dem Datensatz *mtcars.*
  - Über die Clusterzuordnung für jedes Auto werden die Durchschnittswerte pro Cluster bestimmt.

[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (1000 lbs)
[, 7] qsec 1/4 mile time
[, 8] vs Engine (0 = V-shaped, 1 = straight)
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

	avg_cyl	avg.horsepower	avg_drat	avg_weight	avg_qsec	share_vmotor
	2	4	5	6	7	8
Cluster 1	7.9	202.6	3.2	3.9	16.9	0.0
Cluster 2	4.7	97.4	3.9	2.6	18.6	0.4

	share_manual_transmission	avg_gears	avg_carb
	9	10	11
Cluster 1	0.1	3.3	3.3
Cluster 2	0.3	4.1	2.4

Interpretation der Clusterlösung

- Erkenntnisse:
  - Cluster 1 hat generell eine höhere Leistung und einen höheren Verbrauch als Cluster 2.
  - Cluster 1-Autos haben durchschnittlich knapp 8 Zylinder, während im Cluster 2 sich typischerweise Autos mit 4 bis 6 Zylindern befinden.
- Mit den gesammelten Informationen interpretieren wir die zwei gefundenen Cluster als zwei Preisklassen.
  - *Das Cluster 1 enthält die luxuriöseren Autos, mit mehr Leistung, mehr Zylindern und höherem Verbrauch, man könnte die Gruppe als 'Fahrzeuge der Oberklasse' betiteln.*
  - *Das Cluster 2 enthält demnach leistungsschwächere Autos, die günstiger im Preis sind und einen geringeren Verbrauch aufweisen.*

[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (1000 lbs)
[, 7] qsec 1/4 mile time
[, 8] vs Engine (0 = V-shaped, 1 = straight)
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

	avg_cyl	avg.horsepower	avg_drat	avg_weight	avg_qsec	share_vmotor
	2	4	5	6	7	8
Cluster 1	7.9	202.6	3.2	3.9	16.9	0.0
Cluster 2	4.7	97.4	3.9	2.6	18.6	0.4

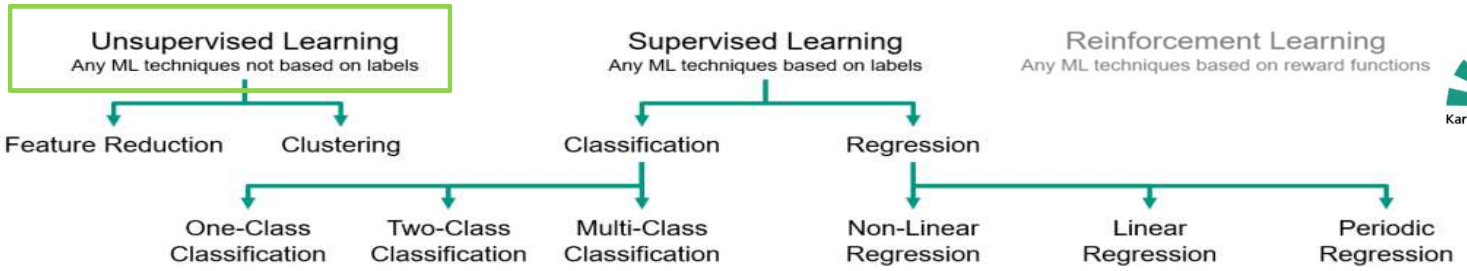
	share_manual_transmission	avg_gears	avg_carb
	9	10	11
Cluster 1	0.1	3.3	3.3
Cluster 2	0.3	4.1	2.4

Interpretation der Clusterlösung

- Um diese Interpretation abzusichern, haben wir die unverbindliche Preisempfehlung (Achtung: die Modelle stammen aus den 1970ern) für die Autos herausgesucht.
- In der unteren Tabelle sind die Durchschnittswerte für jedes der beiden Cluster angegeben, in Klammern steht zusätzlich der Median.

Durchschnittl. Preisempfehlung Cluster 1	Durchschnittl. Preisempfehlung Cluster 2
9974 Dollar (7474)	6283 Dollar (4295)

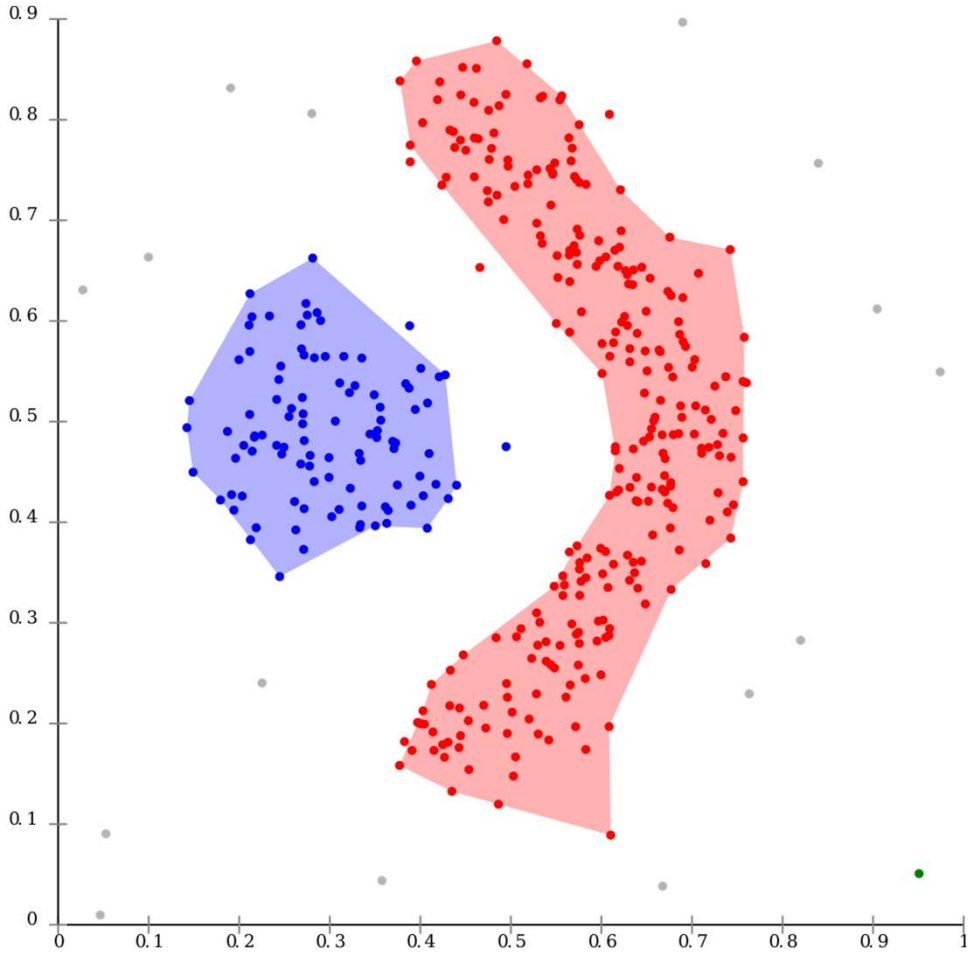
Machinelles Lernen



Algorithm ↓	Abk.	Feature Reduction	Clustering	One-Class Classification	Two-Class Classification	Multi-Class Classification	Non-Linear Regression	Linear Regression	Periodic Regression
Faktorenanalyse		X							
Principal Component Analysis	PCA	X		X					
	K-means		X						
hierarchical cluster analysis	HCA		X						
DBSCAN			X						
One Class Support Vector Machine	OCSVM			X					
Isolation Forest				X					
	LODA			X					
(künstliche) Neuronale Netze	NN			X (Autoencoder)	X	X	X	X	X
Support Vector Machine	SVM				X				
Decision Tree					X	X			
Bayes-Klassifikation					X	X			
Random Forest						X			
Diskriminanzanalyse				x	x	x			
Logistic Regression							X	X	
Linear Regression								X	
Harmonic Regression									X
Nächste-Nachbar-Klassifikation	K-NN	x	x						

Dichtebasierte Verfahren

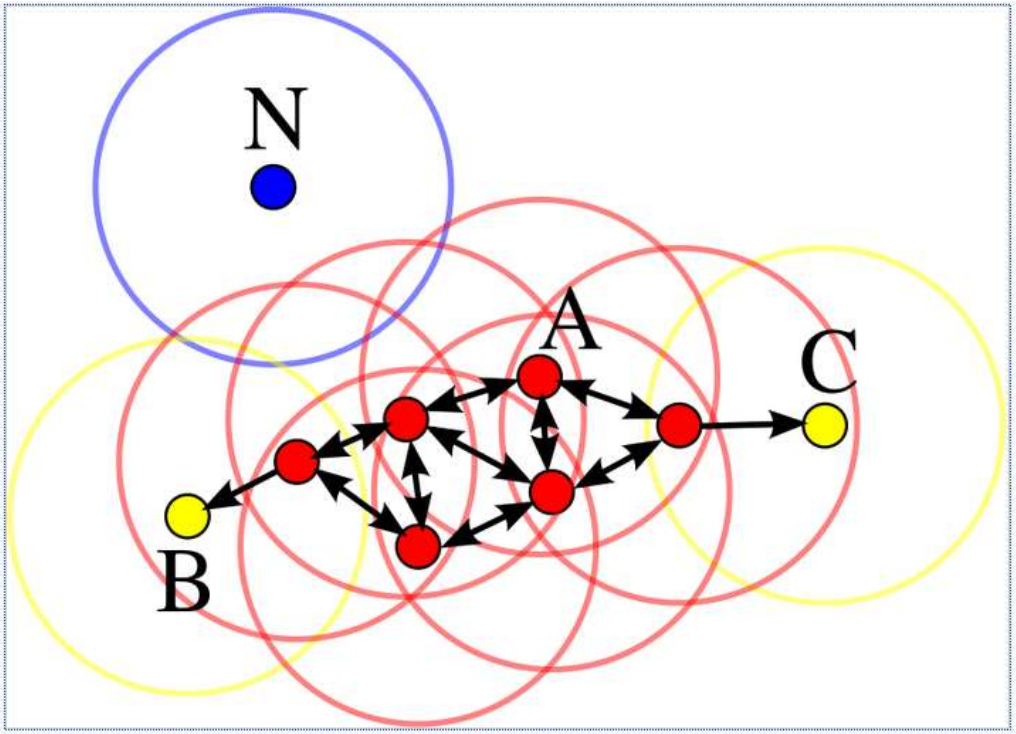
- Bei dichtebasiertem Clustering werden Cluster als Objekte in einem mehr-dimensionalen Raum betrachtet, welche dicht beieinander liegen, getrennt durch Gebiete mit geringerer Dichte.
- Dichtebasierte Verfahren
  - Dichtebasierte Clusterverfahren wurden entwickelt, um Cluster mit unregelmäßigen Formen darstellen zu können.
  - Diese Methoden erkennen Cluster als Regionen mit einer hohen Dichte an Objekten.
  - Die Cluster werden durch Regionen mit geringer Objektdichte voneinander getrennt.
  - Gruppen werden nach der Dichte der Punktewolke gruppiert.



DBSCAN

Dichtebasiertes Verfahren

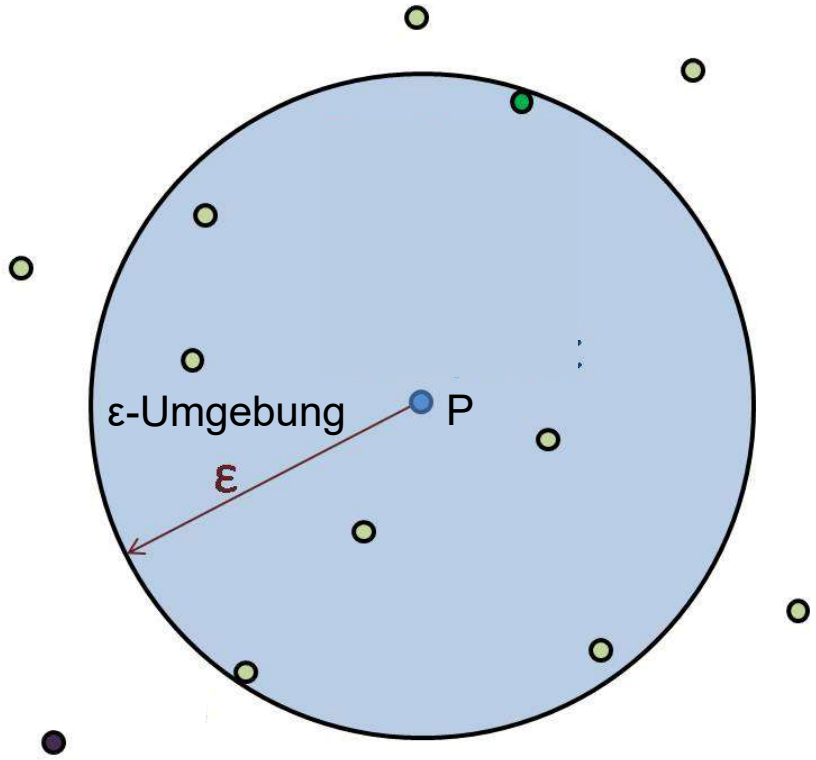
- Die bekannteste dichtebasierte Methode ist DBSCAN
  - *Density Based Spatial Clustering of Applications with Noise*, 1996 von Ester et al. veröffentlicht.
- DBSCAN clustert folgendermaßen:
  - Für jedes Objekt wird eine Nachbarschaft (konzentrischer kreis mit Radius ε) festgelegt.
  - Wenn sich innerhalb dieser Nachbarschaft mindestens eine definierte Anzahl von Objekten *minPts* befindet, dann wird dieses Objekt ein Kern genannt.
  - Die Objekte in der Nachbarschaft eines Kernes können jedoch selbst wieder Kerne sein.
  - Kerne, die sich in der Nachbarschaft eines anderen Kerns befinden werden mit diesem verknüpft.
    - ► Solche Verbindungen werden als Region hoher Objektdichte oder als Cluster bezeichnet.
  - Bedingt durch diese Form des Clusterwachstums können die Cluster unregelmäßige Formen annehmen.
  - Alle Objekte, die schließlich nicht in einem Cluster enthalten sind werden als Störgeräusche aufgefasst.



Punkte bei A sind Kernpunkte. Punkte B und C sind *dichte-erreichbar* von A und dadurch *dichte-verbunden* und gehören zu dem gleichen Cluster. Punkt N ist weder ein Kernpunkt, noch dichte-erreichbar, also Rauschen. ($minPts$=3 oder $minPts$=4)

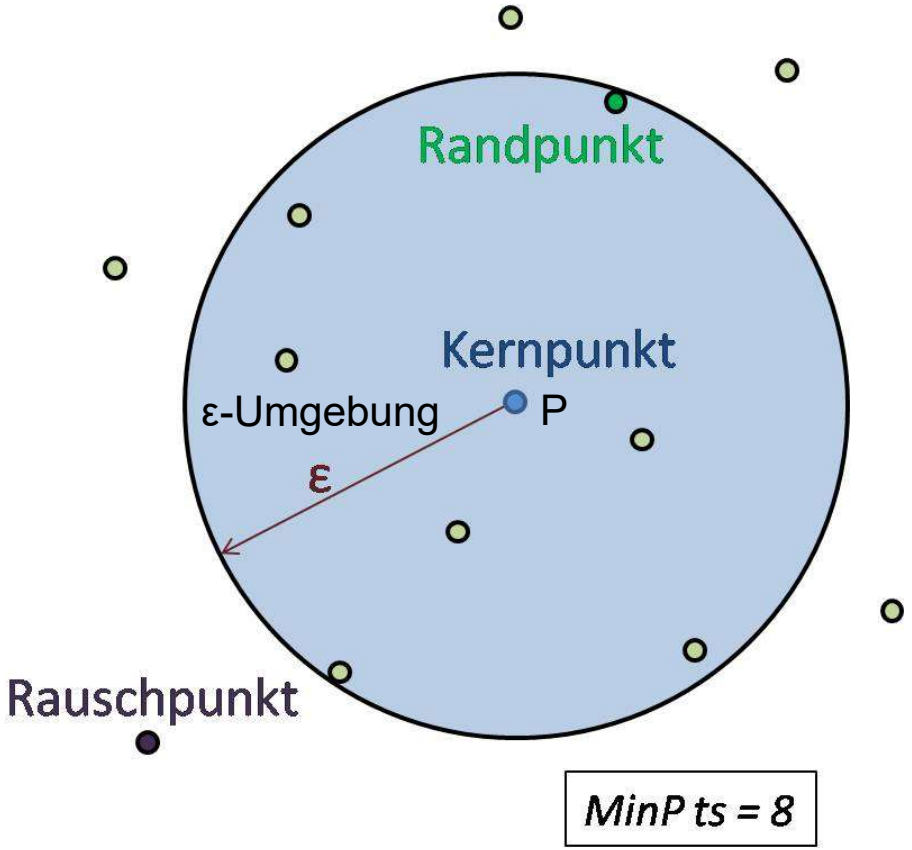
DBSCAN
Vorgehensweise

- Zunächst muss man sich überlegen, wie die Dichte eines speziellen Datenpunkts bemessen werden kann.
  - Beim zentrumsbasierten Zugang wird die Dichte für den Punkt p durch die Anzahl der Punkte in einer ε-Umgebung für ein zu bestimmendes ε geschätzt.
  - Die ε-Umgebung von p besteht aus allen Punkten, die höchstens den Abstand ε zu p besitzen.
- Ist der Radius ε sehr groß, so besitzen alle Datenpunkte die Dichte m (= Anzahl Datenpunkte), da dann für jeden Punkt p alle Punkte in der ε-Umgebung von p liegen.
  - Wählt man ε zu klein, so hat jeder Punkt die Dichte 1.
    - ► Daher ist die Bestimmung von ε im Vorfeld der Analyse von zentraler Bedeutung.
  - Ein weiterer Parameter, den wir für den DBSCAN-Algorithmus benötigen, ist MinPts.
    - ► Dieser legt fest, wie viele Punkte in der ε-Umgebung von einem Punkt p liegen müssen, damit p zu einem Cluster gehört.



Daten in drei verschiedene Kategorien

- Ein Datenpunkt liegt entweder im Inneren einer dichten Region, auf dem Rand einer solchen oder in einem spärlich besetzten Gebiet.
- Kernpunkt:
  - Die Anzahl der Datenpunkte in der ε-Umgebung des Kernpunkts beträgt mindestens MinPts.
- Randpunkt:
  - Ein Randpunkt ist kein Kernpunkt, liegt aber in der ε-Abstand eines Kernpunktes.
- Rauschpunkt:
  - Ein Rauschpunkt ist weder Kern- noch Randpunkt.



DBSCAN

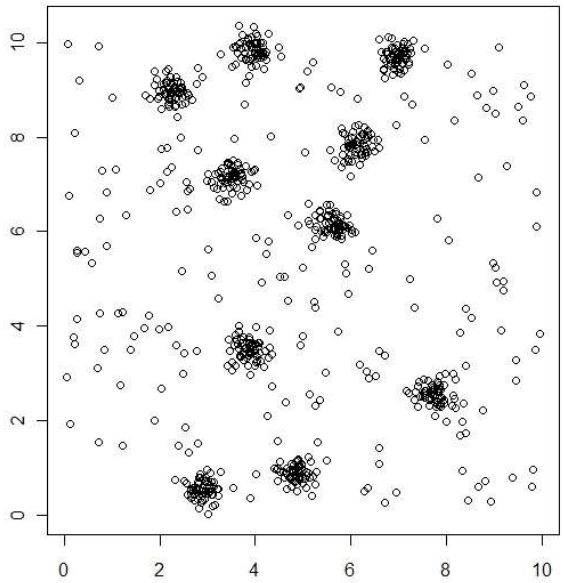
Algorithmus <zur persönlichen Vertiefung>

- Informell:
  - Es landen zwei Kernpunkte, deren Abstand höchstens ε beträgt, im gleichen Cluster, Randpunkte werden dem Cluster eines entsprechenden Kernpunktes zugeordnet und Rauschpunkte werden vernachlässigt.
- Zeit- und Speicherplatz-Aufwand
  - Der Zeitaufwand beträgt $O(m \times t)$, wobei t die Zeit zum Zählen der Punkte in einer ε-Umgebung ist, im Worst-Case daher $O(m^2)$.
    - Das bedeutet, im schlimmsten Fall benötigt man m^2 viele elementare Rechnungen, wobei m die Anzahl der Datenpunkte ist.
- Auswahl der DBSCAN-Parameter → Damit DBSCAN ein „gutes“ Clustering liefert, muss man die Werte der Parameter ε und MinPts „geschickt“ wählen.
  - Der übliche Ansatz zur Bestimmung dieser Werte, ist die Betrachtung der Distanz eines Punktes p zu seinem k.-nächstem Nachbarn, also dem Punkt, der den k-größten Abstand zu p hat.
  - Diese Distanz werden wir mit k-dist bezeichnen.
  - Für Punkte innerhalb eines Clusters ist k-dist in der Regel klein, für Rauschpunkte dagegen vergleichsweise groß.
  - Wir berechnen daher k-dist für alle Datenpunkte und zeichnen das Ergebnis aufsteigend in ein Diagramm.
  - Man sollte dann einen scharfen Übergang erkennen können, welcher eine gute Wahl für ε nahelegt. k sollte in etwa die Größe der zu erwartenden Cluster widerspiegeln, für zweidimensionale bietet sich k = 4 an.
  - MinPts geben wir schließlich den Wert von k.

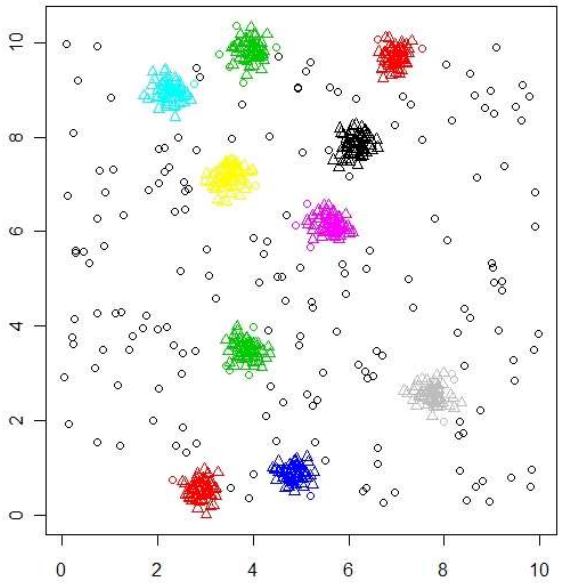
DBSCAN

Vor- und Nachteile von DBSCAN / Beispiele

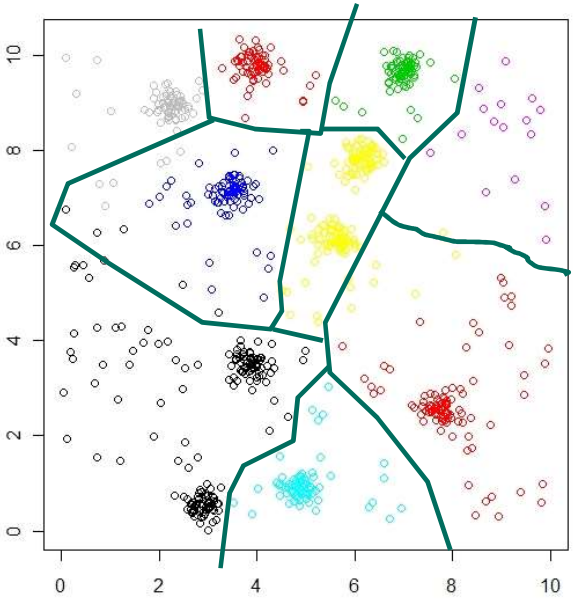
- DBSCAN filtert Rauschen meist sehr gut aus den Daten und kann Cluster beliebiger Form und Größe erfassen.
- Beispiel: Zunächst ein verrauschter Datensatz mit 10 natürlichen Clustern.
  - DBSCAN erkennt diese problemlos, wohingegen der K-Means-Algorithmus kein erwünschtes Clustering liefert; das Rauschen wird nicht unterdrückt.



Datensatz mit 10 Clustern und starkem Rauschen



Cluster in DBSCAN

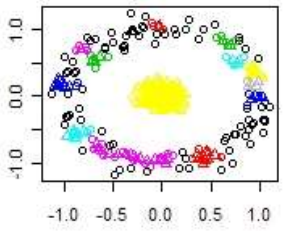


Cluster in K-Means

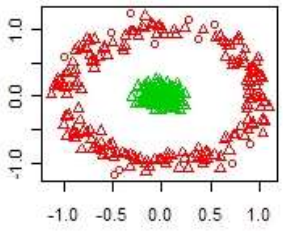
DBSCAN

Beispiel: Datensatz, welcher aus einem Ring und Häufungspunkt in der Ringmitte besteht.

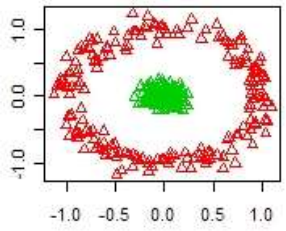
- Bei einem Standartwert *MinPts = 5* kann man ε zwischen 0,2 und 0,6 frei wählen, um das natürliche Clustering zu erhalten (s. rechts)
- Auch bei der Wahl von *MinPts* muss man deutlich vom Standartwert abweichen, bis DBSCAN ein unerwünschtes Clustering wählt.



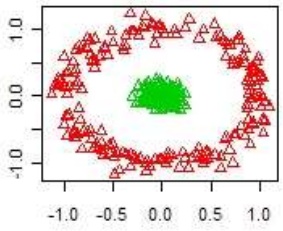
fail



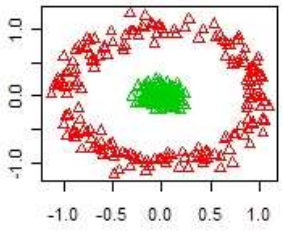
ok



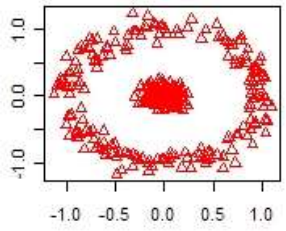
ok



ok

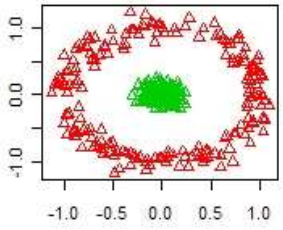


ok

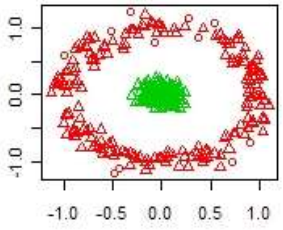


fail

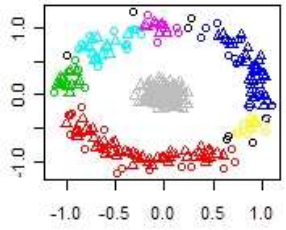
DBSCAN



ok



ok



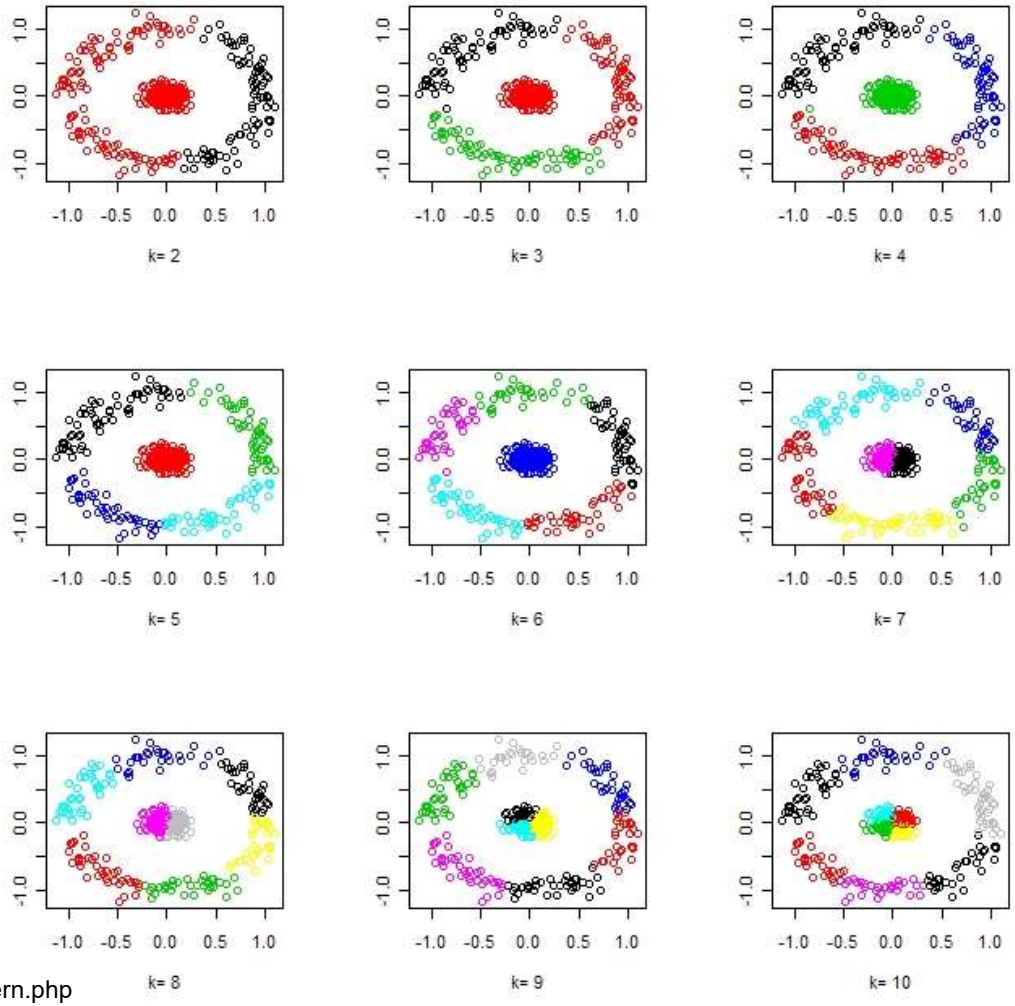
fail

Quelle: http://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/Dichteverbundenes_Clustern.php

DBSCAN

Beispiel: Datensatz, welcher aus einem Ring und Häufungspunkt in der Ringmitte besteht.

- Bei einem Standartwert *MinPts = 5* kann man ε zwischen 0,2 und 0,6 frei wählen, um das natürliche Clustering zu erhalten (s. rechts).
- Auch bei der Wahl von *MinPts* muss man deutlich vom Standartwert abweichen, bis DBSCAN ein unerwünschtes Clustering wählt.
- DBSCAN behandelt Daten mit Clustern besonderer Form recht gut, wohingegen K-Means beim selben Datensatz erhebliche Probleme bekommt (s. rechts)

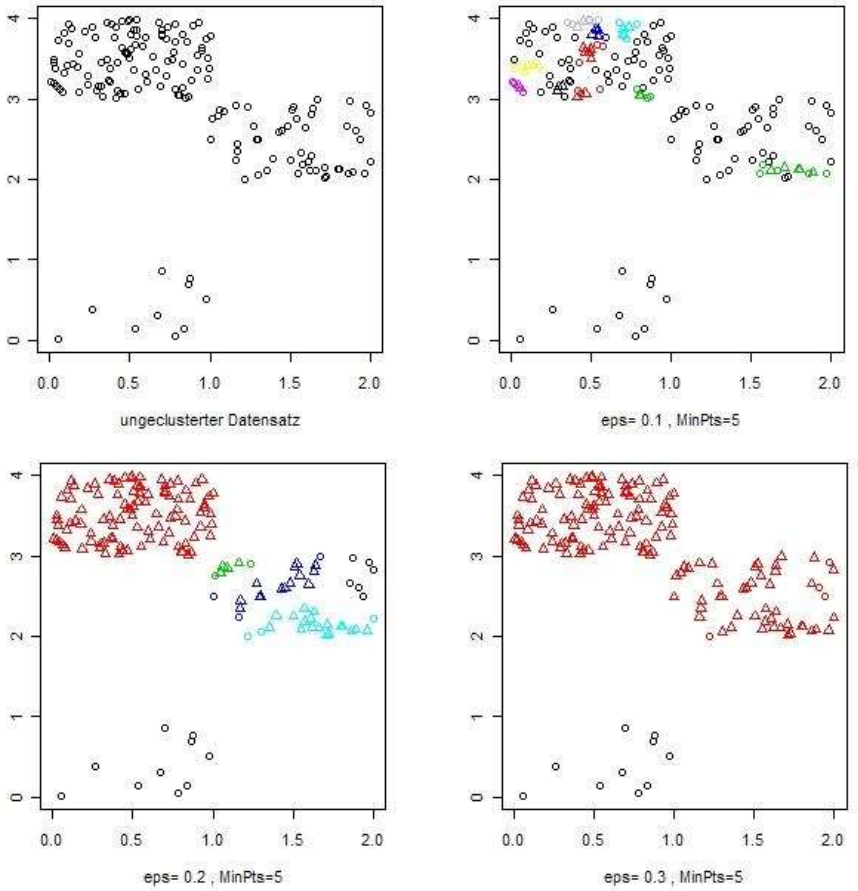


k-Means

Quelle: http://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/Dichteverbundenes_Clustern.php

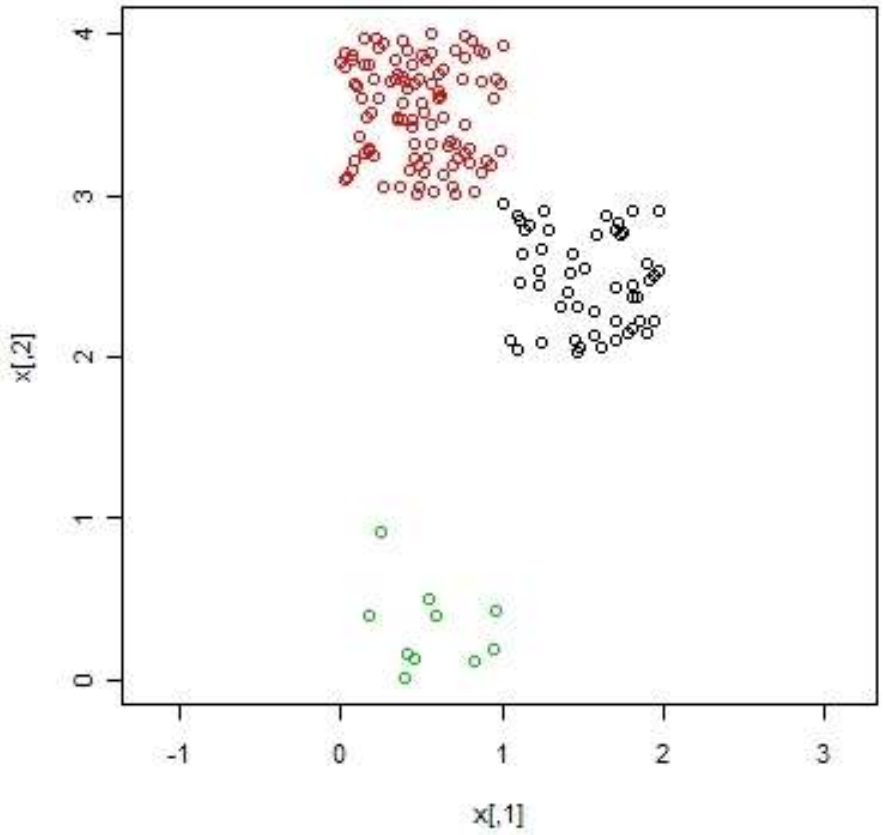
Schwierigkeiten des DBSCAN

- Andererseits hat der DBSCAN-Algorithmus Schwierigkeiten mit der Erkennung von Clustern unterschiedlicher Dichten.
  - Wählt man ε sehr groß, um „schwache“ Cluster zu finden, werden möglicherweise Rauschgebiete als Cluster erkannt.
  - Im rechten Beispiel findet man keine Parameterwerte für ε und MinPts, sodass DBSCAN die drei natürlichen Cluster (Rechtecke) erkennt, da die Punktdichten in den Objekten sehr unterschiedlich sind.



Schwierigkeiten des DBSCAN

- Andererseits hat der DBSCAN-Algorithmus Schwierigkeiten mit der Erkennung von Clustern unterschiedlicher Dichten.
  - Wählt man ε sehr groß, um „schwache“ Cluster zu finden, werden möglicherweise Rauschgebiete als Cluster erkannt.
  - Im rechten Beispiel findet man keine Parameterwerte für ε und MinPts, sodass DBSCAN die drei natürlichen Cluster (Rechtecke) erkennt, da die Punktdichten in den Objekten sehr unterschiedlich sind.
- K-Means erkennt die Cluster hingegen problemlos.
- Außerdem treten beim DBSCAN-Algorithmus Probleme bei hochdimensionalen Daten (Datensatz mit vielen Attributen) auf, da hier die Dichte schwierig zu definieren ist.
- Schließlich kann DBSCAN teuer werden, wenn die Berechnung der nächsten Nachbarn aufwändig ist.



Dichtebasierte Verfahren

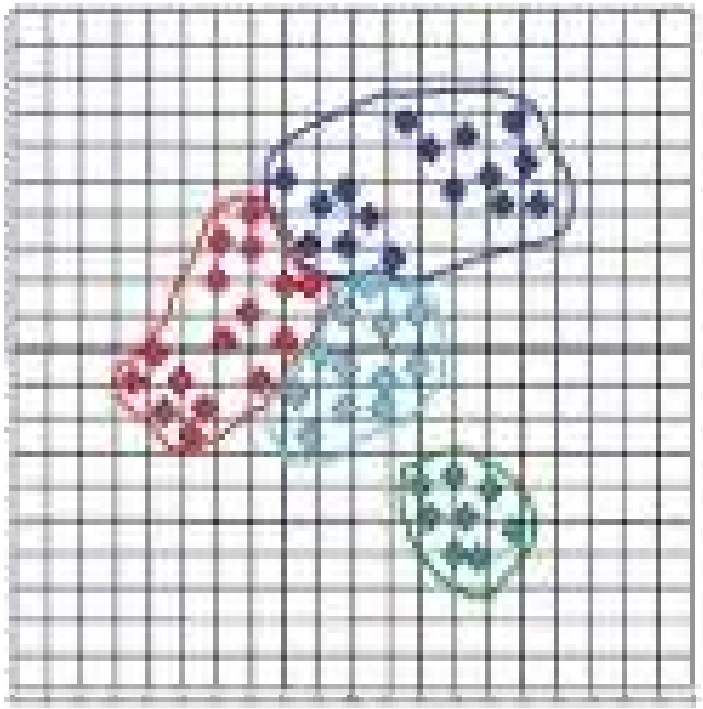
Weitere Beispiele <zur individuellen Vertiefung>

- OPTICS (Ordering Points To Identify the Clustering Structure)
  - Der Algorithmus erweitert DBSCAN, so dass auch verschieden dichte Cluster erkannt werden.
  - Die Wahl des Parameters ε ist nicht mehr so ausschlaggebend um die Clusterstruktur der Objekte zu finden.
- Maximum-Margin Clustering
  - Es werden (leere) Bereiche im Raum der Objekte gesucht, die zwischen zwei Clustern liegen. Daraus werden Clustergrenzen bestimmt und damit auch die Cluster.
  - Die Technik ist eng angebunden an Support-Vektor-Maschinen.

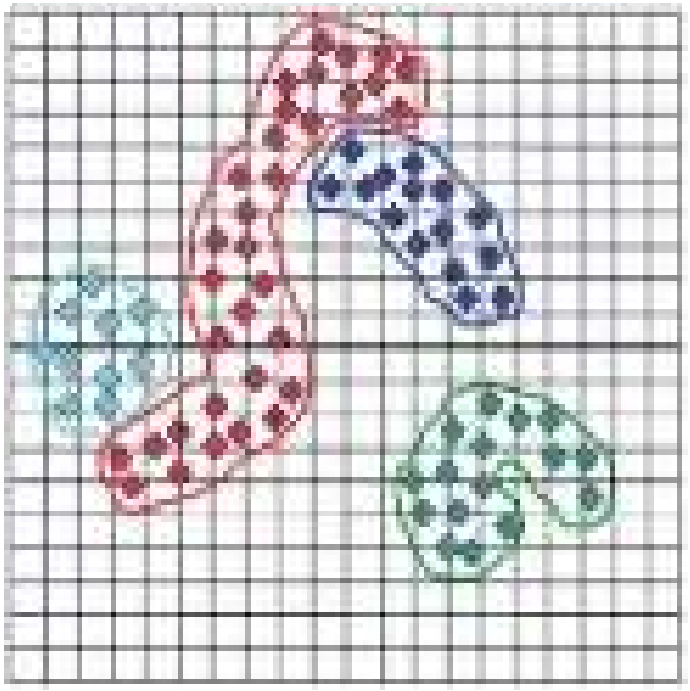
Vergleich: dichtebasiertes und hierarchisches Clusterverfahren

Beispiel: Dichtebasiert

- Beispiel 1



- Beispiel 2

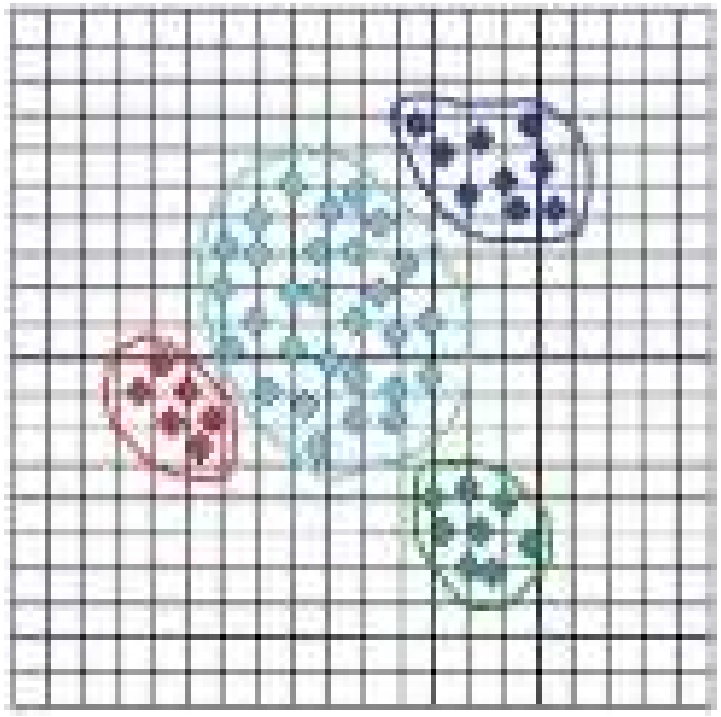


- Dichtebasierte Verfahren erkennen langgezogene Cluster oder Cluster mit unregelmäßigen Formen gut, wohingegen hierarchische Verfahren zu einer eher kugelförmige Clusterbildung tendieren.

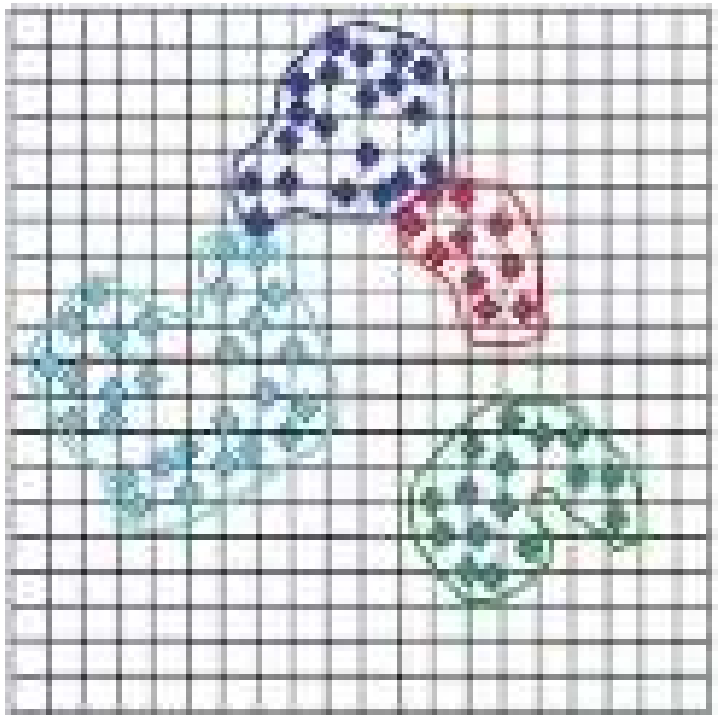
Vergleich: dichtebasiertes und hierarchisches Clusterverfahren

Beispiel: hierarchisch

- Beispiel 1



- Beispiel 2



- Dichtebasierte Verfahren erkennen langgezogene Cluster oder Cluster mit unregelmäßigen Formen gut, wohingegen hierarchische Verfahren zu einer eher kugelförmige Clusterbildung tendieren.

Gitterbasierte Verfahren

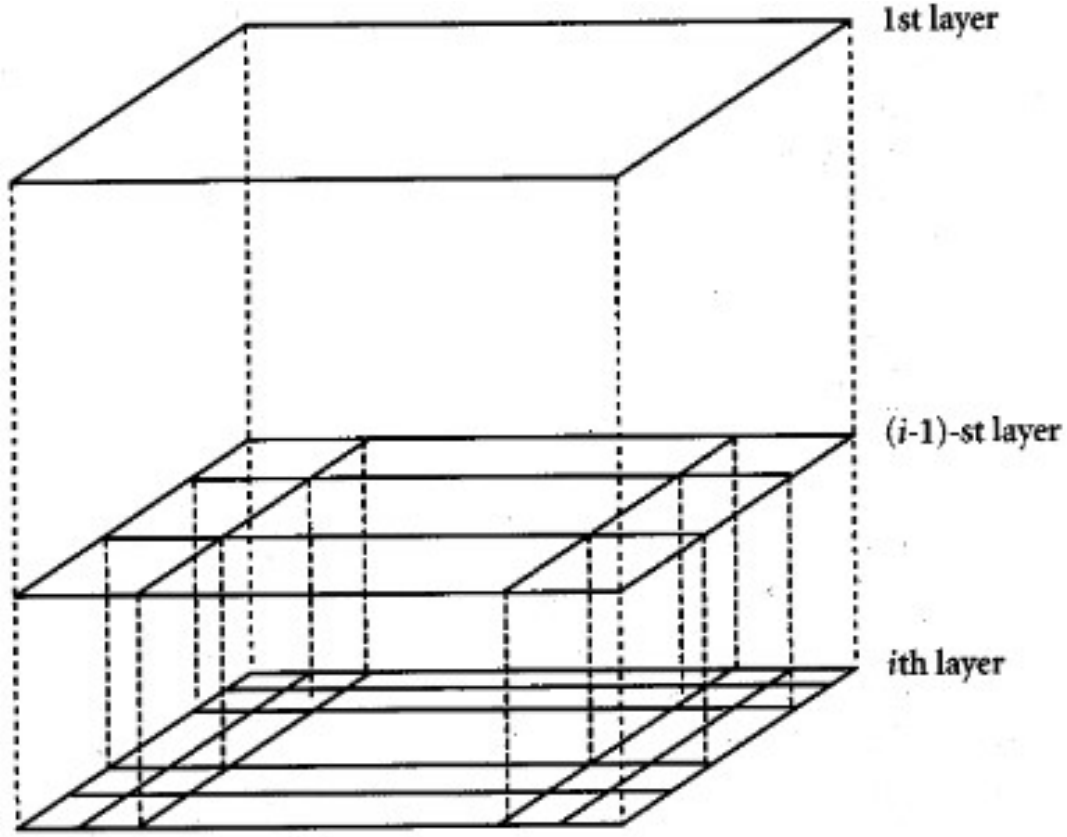
<für interessierte>

- Gitterbasierte Verfahren unterteilen den gesamten Raum in eine endliche Anzahl von Zellen, welche eine Gitterstruktur formen, auf der alle Clusteroperationen durchgeführt werden.
  - Da die Objekte im Raum meist nicht gleichmäßig verteilt sind existieren in der Regel Zellen mit einer höheren Dichte an Objekten.
  - Sie sind somit insbesondere für große und hochdimensionale Suchräume geeignet.
- Eine weitere Strategie bei der Suche nach dicht besetzten Gebieten ist, die Dimension des Suchraums zu verkleinern.
  - Wird in diesem niedriger dimensionalen Raum keine Häufung von Objekten gefunden, so ist dies auch in höheren Dimensionen nicht zu erwarten.
- Aufgrund der rechteckigen Form der Zellen können die Cluster jedoch bei vielen Methoden nur horizontal oder vertikal abgegrenzt werden.
  - Um differenzierte, unregelmäßige Cluster darstellen zu können, muss die Gitterstruktur entsprechend fein gewählt werden. Dies vermindert jedoch die Effizienz der Algorithmen.
- Gitterbasierte Methoden sind beispielsweise STING (Statistical Information Grid), WaveCluster oder CLIQUE (Clustering in Quest).

Gitterbasierte Verfahren

<für interessierte>

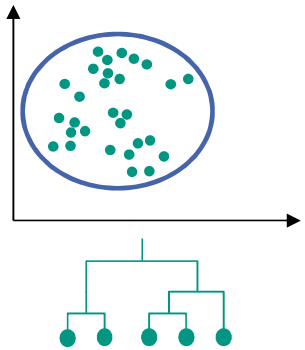
- Ausgehend von der detailliertesten Ebene werden für alle Zellen statistische Maße (z.B. Anzahl, Mittelwert, Standard-abweichung, Min, Max, Verteilungstyp) berechnet, die nach „oben“ aggegriert werden.
  - Cluster können durch Zusammenfassung einzelner Zellen wachsen.
- Vorteil:
  - Effizienz (nur abhängig von #Zellen, nicht von #Objekten)
- Nachteil:
  - Clustergrenzen können nur horizontal oder vertikal sein.



Modellierung der Daten – Data Mining

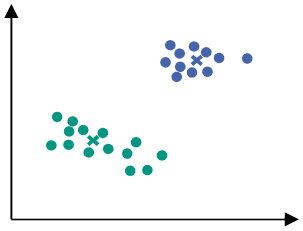
Clustering – Verschiedene Verfahren

Hierarchisch



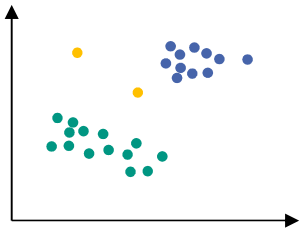
- Verschmelzen des geringsten Abstands
- Bis alle im gleichen Cluster

Partitionierend



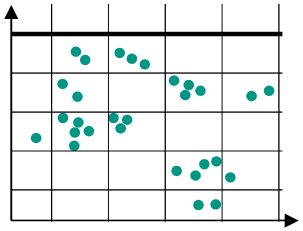
- Anzahl Cluster vorgegeben
- Random Seed
- Zugehörigkeit bestimmen
- Clustermittelpunkte neu berechnen
- Iterationsanzahl oder bis sich nichts mehr ändert

Dichte-basiert



- Cluster, Nachbarn und Dichte gegeben
- Berechnung ob Kern, Nachbar oder Rauschen
- Nachbar von Kernen ausweiten
- Iterationsanzahl oder bis alle Punkte durchlaufen sind

Gitter-basiert

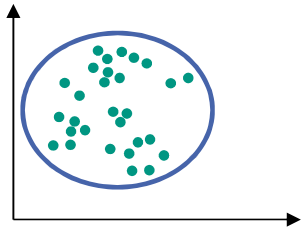


- Dichte der Gitter, Punkte
- Berechnung Punkte/Gitter
- Punkte > e

Modellierung der Daten – Data Mining

Clustering – Verschiedene Verfahren

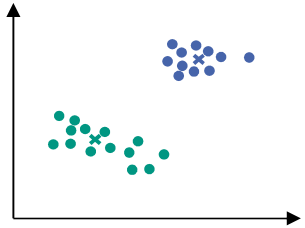
Hierarchisch



\+
- Cluster Hierarchie
- Keine Parameterwahl

-
- Keine Iteration
- Keine Optimierung

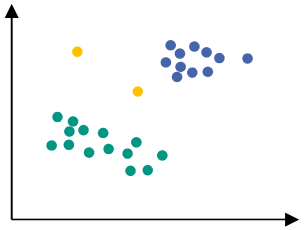
Partitionierend



\+
- Schnell
- Daten unabhängig

-
- Anzahl Cluster muss gewählt werden

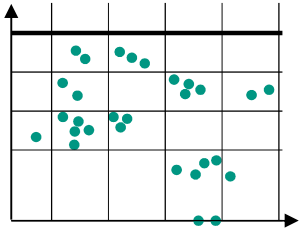
Dichte-basiert



\+
- Unregelmäßige Formen
- Rauschunterdrückung

-
- Sensibel auf Parameterwahl

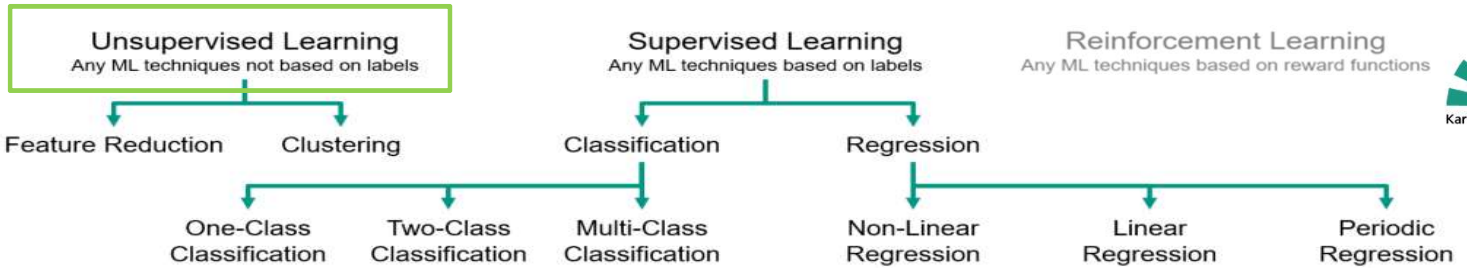
Gitter-basiert



\+
- Für Hochdimensionale Räume geeignet

-
- Eher grob
- Weniger Optimierungspotenzial

Machinelles Lernen



Algorithm ↓	Abk.	Feature Reduction	Clustering	One-Class Classification	Two-Class Classification	Multi-Class Classification	Non-Linear Regression	Linear Regression	Periodic Regression
Faktorenanalyse		X							
Principal Component Analysis	PCA	X		X					
	K-means		X						
hierarchical cluster analysis	HCA		X						
DBSCAN			X						
One Class Support Vector Machine	OCSVM			X					
Isolation Forest				X					
	LODA			X					
(künstliche) Neuronale Netze	NN			X (Autoencoder)	X	X	X	X	X
Support Vector Machine	SVM				X				
Decision Tree					X	X			
Bayes-Klassifikation					X	X			
Random Forest						X			
Diskriminanzanalyse				x	x	x			
Logistic Regression							X	X	
Linear Regression								X	
Harmonic Regression									X
Nächste-Nachbar-Klassifikation	K-NN	x	x						

Hauptkomponentenanalyse (s. 6.2a)

Konstruktive Datenaufbereitung
Data Preparation

- Die Hauptkomponentenanalyse (Principal Component Analysis – PCA) ist eine variablenorientierte Methode, mit der versucht wird, einen hochdimensionalen Datensatz in einen niederdimensionalen Raum zu projizieren.

- Die zentrale Idee dabei ist, die Daten so auf eine 2-dimensionale Ebene zu projizieren, dass die gesuchten Zusammenhänge sichtbar werden.
  - Die sichtbar werdende Struktur der projizierten Daten hängt von der Richtung der Projektion ab.
  - Es stellt sich die Frage, wie eine Rotation der Daten (oder der Achsen - was für diese Zwecke dasselbe ist) gefunden werden kann, die ein Maximum an Information im projizierten Bild darstellt.

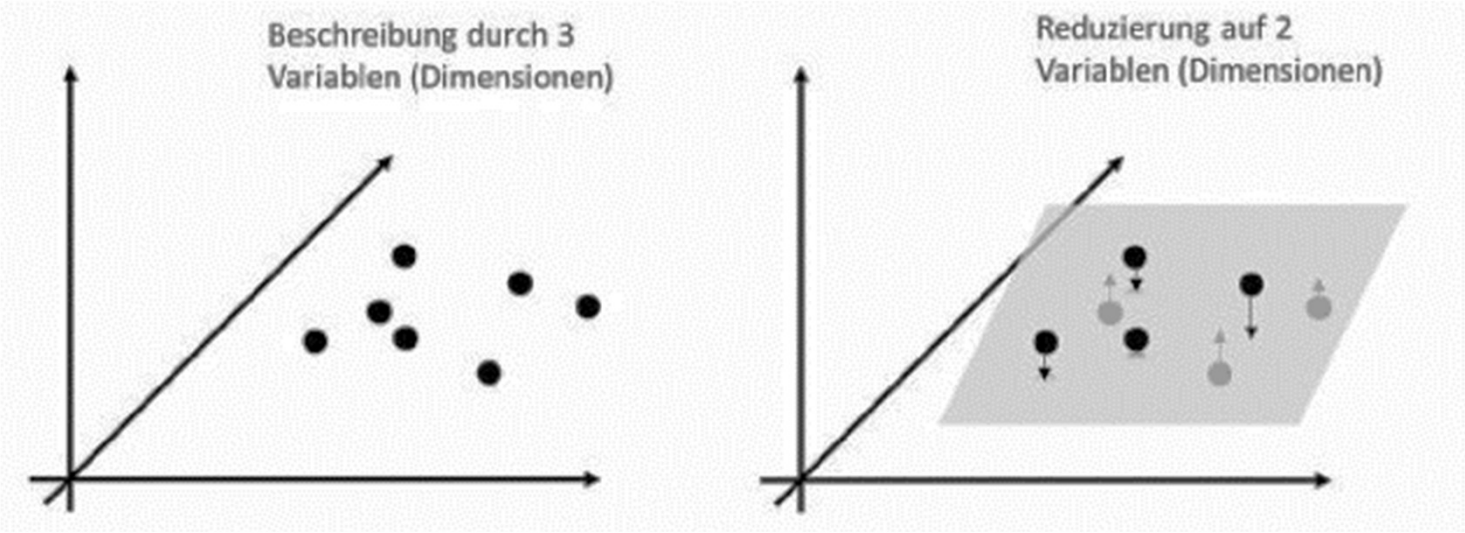
Faktorenanalyse

- Bei der Faktorenanalyse handelt es sich um ein Verfahren zur Zusammenfassung von Variablen.
  - Ziel ist es, die Anzahl der erklärenden Variablen zu verringern und „ähnliche“ Variablen zu Faktoren zusammenzufassen, die dann weitgehend voneinander unabhängig sind.
  - Die Bedeutung der Faktoren ist dabei nicht immer offensichtlich und muss interpretiert werden.
- Die Faktorenanalyse findet also dann Anwendung, wenn eine Vielzahl von Variablen vorhanden sind und davon ausgegangen werden kann, dass diese Variablen oft das Gleiche oder Ähnliches „aussagen“ und sich deshalb auf eine deutlich kleinere Anzahl aussagekräftiger Faktoren reduzieren lassen.
- Ein einfaches Beispiel hierzu bildet die Verdichtung der zahlreichen technischen Eigenschaften von PKWs auf wenige Dimensionen, wie z. B. Größe, Leistung, Prestige und Sicherheit.
- Dem Ziel der Faktorenanalyse – nämlich die Reduktion der Variablen und damit die Reduktion der Komplexität des Modells – steht auf der anderen Seite die Gefahr des Informationsverlustes gegenüber.
  - Diese Aspekte müssen gegeneinander abgewogen werden.

Prinzip der Faktorenanalyse

Bildlich kann man das an diesem Beispiel veranschaulichen:

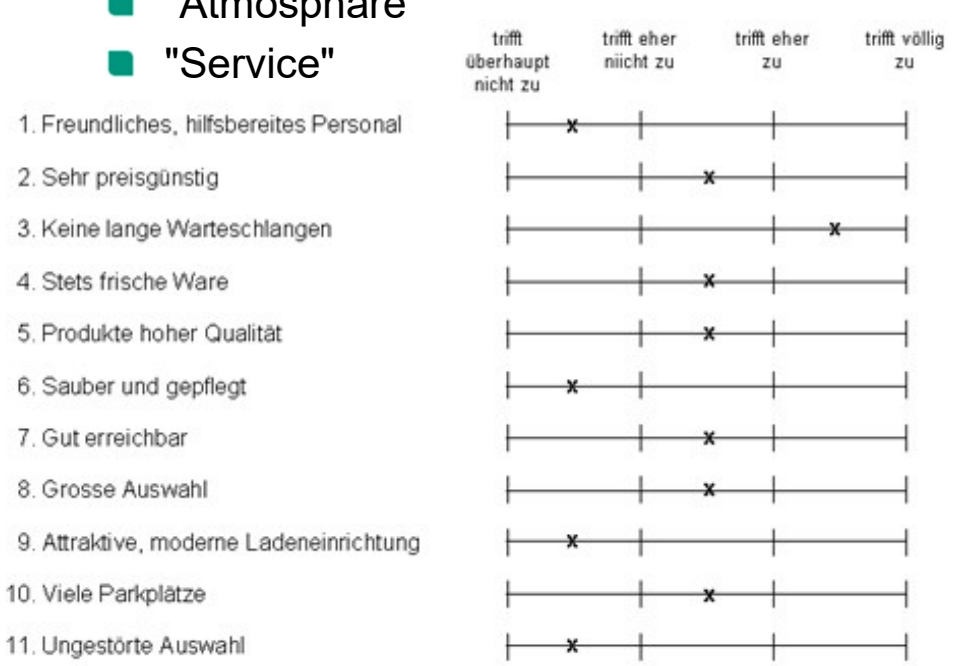
- Mit der Faktorenanalyse wird nun also versucht, eine zweidimensionale Fläche (zwei Faktoren) so in den Raum einzupassen, dass die Punkte mit möglichst wenig Verlust (Minimierung der Pfeillängen) auf dieser Fläche dargestellt werden können.
- Im Ergebnis können die Punkte auf der zweidimensionalen Fläche dargestellt werden, wobei die Reduktion der Komplexität mit einem gewissen Informationsverlust einhergeht.

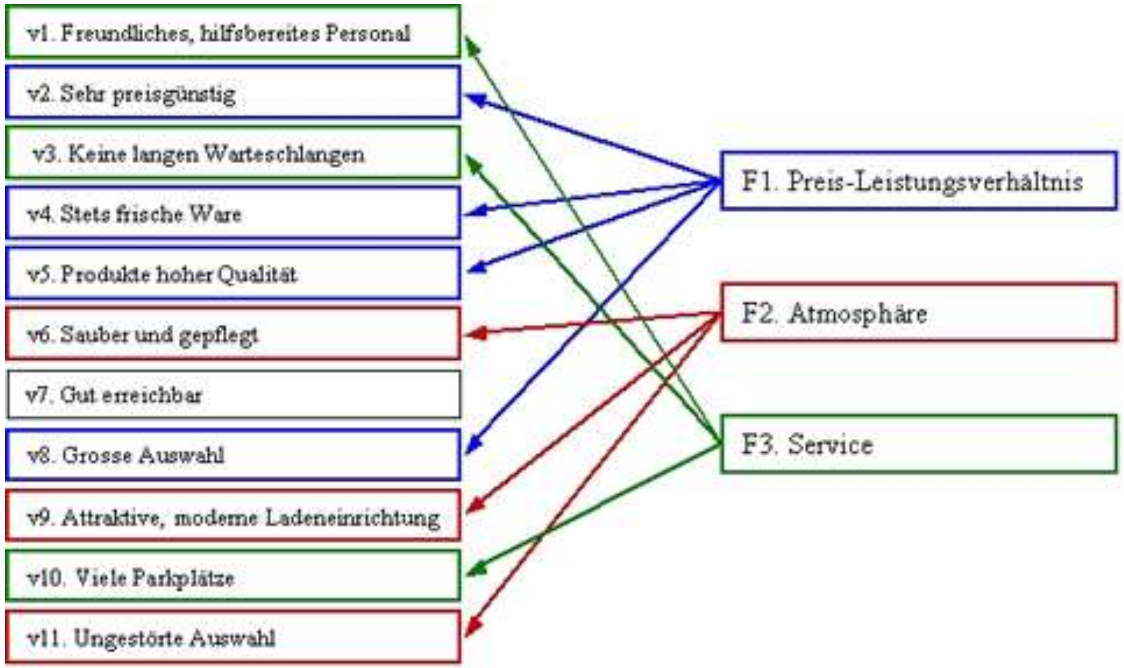


Faktorenanalyse

Illustratives Beispiel

- Bei der Untersuchung zum Image eines Ladengeschäftes wurde die Einschätzung von Kunden bezüglich von elf Eigenschaften zum Image mit einer 4-stufigen Skala abgefragt.
- Ziel: die 11 Items zum Thema Image auf wenige Faktoren, konkret auf folgende 3 zu reduzieren:
  - "Preis-Leistungsverhältnis“
  - "Atmosphäre"
  - "Service"





s. http://www.mri.imh.unisg.ch/analysemethoden/datenanalyse/deskriptiv/multivariat/Faktorenanalyse.html

Hauptkomponentenanalyse vs. Faktorenanalyse

Zusammenfassung

- Die Hauptkomponentenanalyse (PCA) ist eine variablenorientierte Methode, mit der – vergleichbar zur Faktorenanalyse – versucht wird, einen hochdimensionalen Datensatz in einen niederdimensionalen Raum zu projizieren.
- Dabei versucht die PCA, die Varianzen der Objekte im ursprünglichen Raum möglichst gut mit dem neuen niederdimensionalen Raum abzudecken.
- Die PCA besteht darin, eine orthogonale Transformation der ursprünglichen Variablen in eine neue Menge unkorrelierter Variablen, die Hauptkomponenten, vorzunehmen.

- Im Gegensatz dazu sind die Faktoren bei der Faktorenanalyse nicht zwingend orthogonal.
- Die Hauptkomponenten werden nacheinander in absteigender Bedeutung konstruiert.
- Die Hauptkomponenten sind Linearkombinationen der ursprünglichen Variablen.
  - Die erste Hauptkomponente wird so konstruiert, dass sie für den größten Teil der Variation verantwortlich ist.

<zur eigenen Vertiefung>

- Die Faktorenanalyse ist eine Familie multivariater Verfahren, bei denen eine Menge von beobachtbaren (manifesten) Variablen auf wenige zugrunde liegende nicht beobachtbare Variablen zurückgeführt wird, die aus den beobachtbaren Variablen zusammengesetzt sind. Diese nicht beobachteten Variablen werden als Faktoren bezeichnet. Bei der hier vorgestellten explorativen Faktorenanalyse sind keine inhaltlichen Vorannahmen nötig. Es wird lediglich untersucht, inwieweit sich die Zusammenhänge zwischen einer Menge von beobachtbaren Variablen durch wenige Faktoren erklären lassen. Weder die Zahl der Faktoren noch die genaue Zuordnung der manifesten Variablen zu den Faktoren ist bekannt. Im Unterschied dazu müssen bei einer konfirmatorischen Faktorenanalyse genaue Hypothesen über die Zahl der Faktoren und die Zuordnung der manifesten Variablen zu den Faktoren vorliegen. In diesem Beitrag werden mit der Hauptkomponentenanalyse und der explorativen Faktorenanalyse zwei Verfahren vorgestellt, die in ihren Grundannahmen zwar verschieden, in der Anwendung aber austauschbar erscheinen können. Dies zeigt sich bereits an der Verwendung des Begriffes Faktorenanalyse: Einerseits steht er für ein bestimmtes Modell, nämlich das Modell mehrerer gemeinsamer Faktoren, andererseits dient der Begriff Faktorenanalyse aber auch als Sammelbegriff für eine Familie von Verfahren, unter den auch die Hauptkomponentenanalyse fällt. Im Folgenden soll eine geometrisch orientierte Darstellung der Hauptkomponentenanalyse den Einstieg ermöglichen und die konkreten Schritte der Hauptkomponentenanalyse dargestellt werden. Anschließend wird das Modell mehrerer gemeinsamer Faktoren präsentiert und von der Hauptkomponentenanalyse abgegrenzt. Abschnitt 2 enthält die mathematischen Grundlagen und ein Anwendungsbeispiel wird in Abschnitt 3 präsentiert. Abschließend werden in Abschnitt 4 häufige Probleme diskutiert und Handlungsempfehlungen abgeleitet.

Fragen?

Inhalt IT2

6. Big Data & Maschinelles Lernen

- Definitionen, Verwendung (Charakteristik, Risiken, Chancen)
- Motivation und Anwendungsfälle
- Data Science Prozesse:
  - KDD
  - CRISP-DM
- CRISP-DM im Detail
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
- Infrastruktur für Big Data
  - Verfahren zur Datenanalyse
  - Beispiel zu ML-Bibliotheken: Tensorflow
  - Hardware für ML

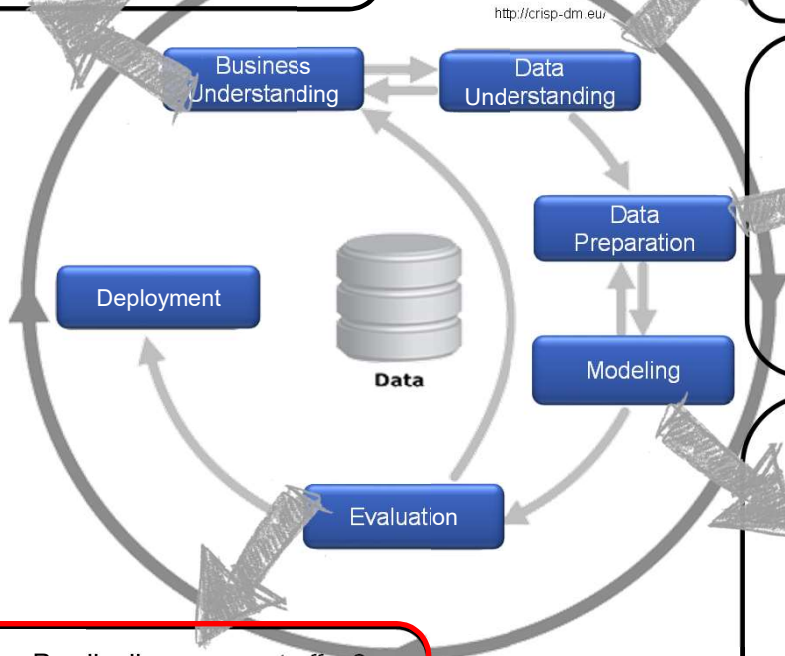
Big Data & Maschinelles Lernen

CRISP-DM im Detail

- Zielsetzung
- Rand-bedingungen

Analytic Aproach

- Regression
- Klassifikation
- Clustering
- Recommendation
- Outlier



Daten
- Struktur
- Statistik
- Visualisierung

Collect Initial Data
- Feature
- Type
- Struktur

Describe & Explore Data
- Statistik: Mean,Var, Median
- Visualisierung: Plots

Daten
Vor- & Aufbe-reitung

Select & Clean Data
- Auswahl
- Konsistenz
- Bereinigen

Construct Data
- Aggregation
- Aufteilung

Format Data
- Syntaktische Modifizierung
- Transformation
- Scaling

Algorithmus Selektion

Modeling Technique
- Group by Learning Style
  - Un- vs Supervised
  - Reinforcement
- Group by Similarity
  - Regression, Neighbor, Clustering, Decision Tree, NN, Bayes

Testdesign
- Splitting
  - Test, Train, Validate

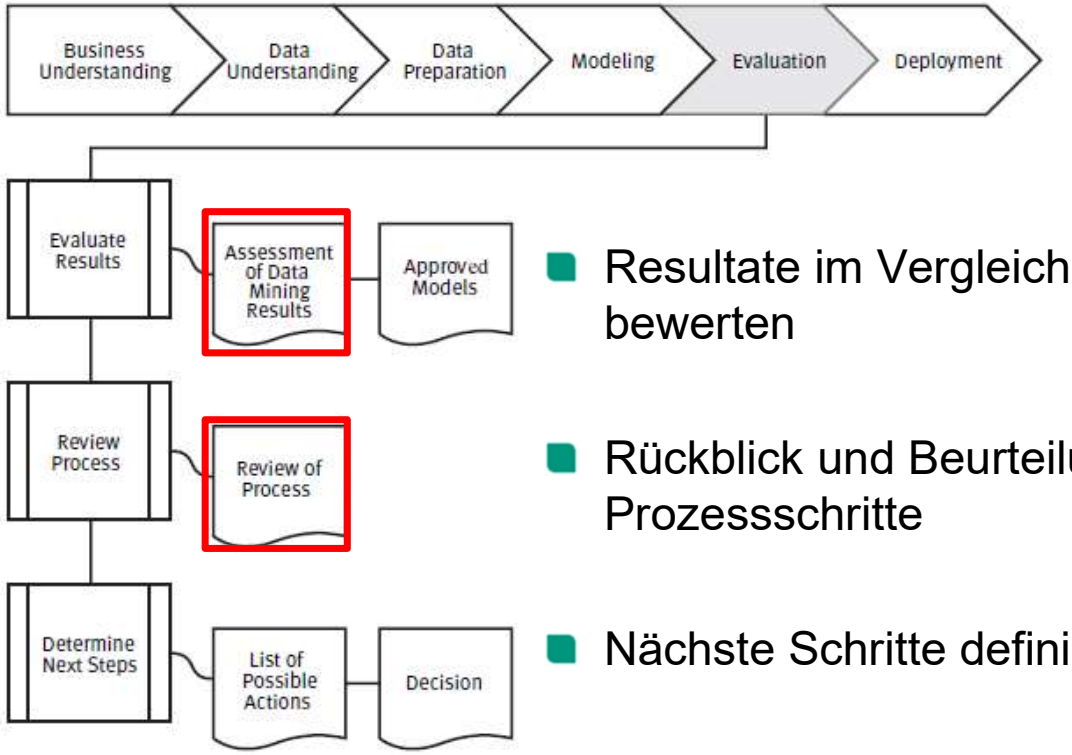
Build & Assess Model
- ROC
- Confusion Matrix
- Cross Validation
- Precision/Recall
- AUC

Ergebnisse

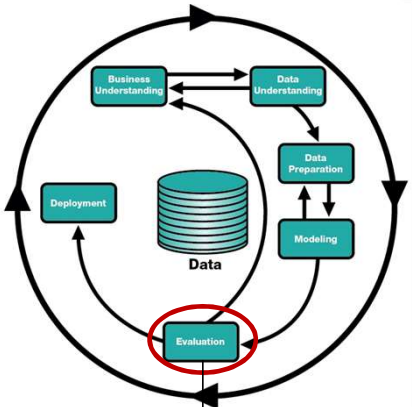
Evaluate Results
- Randbedingungen getroffen? HW Constraints
- Summarize assessment results

Review of process
- Summarize process and highlight activites

CRISP-DM im Detail: Evaluation



- Resultate im Vergleich zum Business-Ziel bewerten
- Rückblick und Beurteilung aller Prozessschritte
- Nächste Schritte definieren



Ergebnisse

Evaluate Results

- Randbedingungen getroffen?
HW Constraints
- Summarize assessment results

Review of process

- Summarize process and highlight activites

Evaluation
(Data Mining-) Resultate auswerten

Resultate = (trainierte) Modelle + neue Erkenntnisse/Wissen

- Resultate verstehen und interpretieren
- Finales Statement:
  - Ursprüngliche Fragestellung gelöst?
  - Gibt es einen geschäftlichen Grund, warum das Modell unzureichend ist?
- Test auf Zielsystem → alle Anforderungen erfüllt? → Deployment möglich?
- Neue Erkenntnisse müssen nicht mit dem (ursprünglichen) Ziel zusammenhängen
  - Bsp: mangelnde Datenqualität festgestellt
  - Neue Fragestellungen / Ziele

Evaluation

Beurteilung aller Prozessschritte, nächste Schritte definieren

- Business-Ziele erreicht, trotzdem Review zur Qualitätssicherstellung nötig
- Alle getroffenen Entscheidungen überprüfen:
  - War es notwendig?
  - Wurde es optimal ausgeführt?
  - Inwiefern könnte es verbessert werden?
  - Beispiel: richtige Merkmalsauswahl, alle Algorithmen betrachtet, passende Daten gesammelt?

Fehler identifizieren und Alternativen für die nächste Iteration / Projekt finden

Karlsruher Institut für Technologie

Fragen?

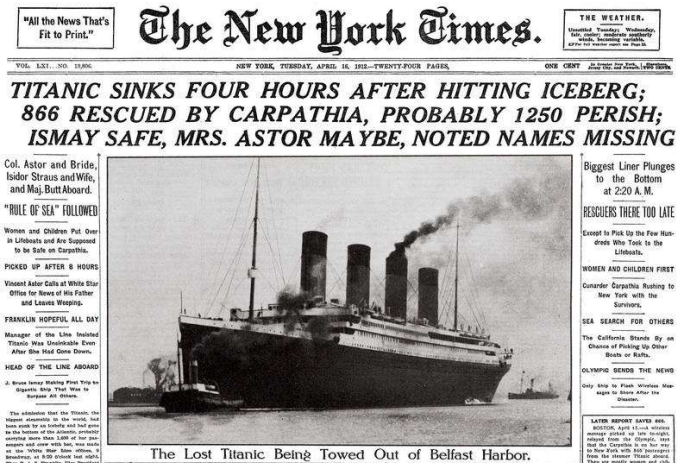
Backup

Beispiele

- Im Haushalt können Neuronale Netze bei Regelungsaufgaben wie der Benutzung einer Waschmaschine oder bei der Steuerung eines Staubsaugerroboters zum Einsatz kommen. Dieser kann mit Hilfe eines solchen Netzes seine Umgebung erlernen und kann bei jedem Saugvorgang seine abgefahrenen Pfade verbessern. Natürlich bleibt Ihre Wohnung nicht vollständig gleich. Hier wird einmal ein Tisch verrückt. Dort bekommt das Sofa einen neuen Standort. Da sich jedoch meist nicht alles gleichzeitig verändern wird, kann die Robotersteuerung durch ein Neuronales Netz die alten Kenntnisse, die sich noch verwenden lassen, weiter behalten und die neuen Kenntnisse zusätzlich lernen. Das heißt, mit Hilfe eines Neuronalen Netzes verblasst seine Erinnerung mit der Zeit und neue Informationen stellen sich in den Vordergrund, so dass er stets das momentan optimale Ergebnis liefert.
- Neuronale Netze können zur Optimierung einer Robotersteuerung eingesetzt werden. Industrieroboter z.B. bei der Produktion von Automobilen haben die Eigenschaft, dass sie aus einer Vielzahl von Gelenken bestehen, mit denen auch ein kompliziertes hineinreichen z.B. in eine Karosserie durchführbar ist. Die Gelenke können jedoch schnell unter Materialermüdung leiden, besonders wenn sie ruckartig bewegt werden und hohe Fliehkräfte wirken. Es geht darum, den Roboterarm möglichst schnell und möglich schonend zu bewegen, so dass alle Arbeiten geleistet werden können. Hierfür können neuronale Netze genutzt werden. Während des Betriebes verändert aufgrund von Abnutzungen der Roboterarm seine Eigenschaften. Eine Korrektur in dem Computerprogramm muss anhand verschiedener Parameter nachgehalten werden. Da sich die Parameter nur langsam ändern, kann auch hierfür ein Neuronales Netz genutzt werden.

Wenn Sie auf der Titanic gewesen wäre…

- Festlegen der Betriebswirtschaftlichen Ziele:
  Wir verkaufen Reiseversicherungen an Passagiere auf Luxusdampfern. Wir wollen wissen
  - Wie hoch ist der monatliche Beitrag in Abhängigkeit der Wahrscheinlichkeit zu sterben.
- Situationsbewertung:
  - Welche Ressourcen haben wir? Personal: Professor Sax, Doktoranden, Studenten. Rechenleistung: Laptop, TitanV-Server
- Bestimmung der Data Mining Ziele:
  - Voraussagung ob die ausgesuchte Person den Untergang der Titanic überlebt hätte
    → Klassifikationsproblem (Überleben ja/nein),
    →gelabelte Daten benötigt
- Erstellung eines Projektplanes:
  - Hier könnte man überlegen wann welcher Teil der Vorlesung stattfindet und anhand dessen einen Projektplan erstellen. Sodass die Evaluation der Frage etwa mit dem wirklichen Monat/Woche in dem die Vorlesung stattfindet übereinstimmt.


"All the News That's Fit to Print."

The New York Times.

THE WEATHER.

NEW YORK, TUESDAY, APRIL 16, 1912.—TWENTY-FOUR PAGES. ONE CENT

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG; 866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH; ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

Women and Children Put Over in Lifeboats and Are Supposed to be Safe on Carpathia.

PICKED UP AFTER 8 HOURS

Vincent Astor Calls at White Star Office for News of His Father and Leaves Weeping.

FRANKLIN HOPEFUL ALL DAY

Manager of the Line Insisted Titanic Was Unsinkable Even After She Had Gone Down.

HEAD OF THE LINE ABOARD

Biggest Liner Plunges to the Bottom at 2:20 A.M.

RESCUERS THERE TOO LATE

Except to Pick Up the Few Hundreds Who Took to the Lifeboats.

WOMEN AND CHILDREN FIRST

Cunarder Carpathia Rushing to New York with the Survivors.

SEA SEARCH FOR OTHERS

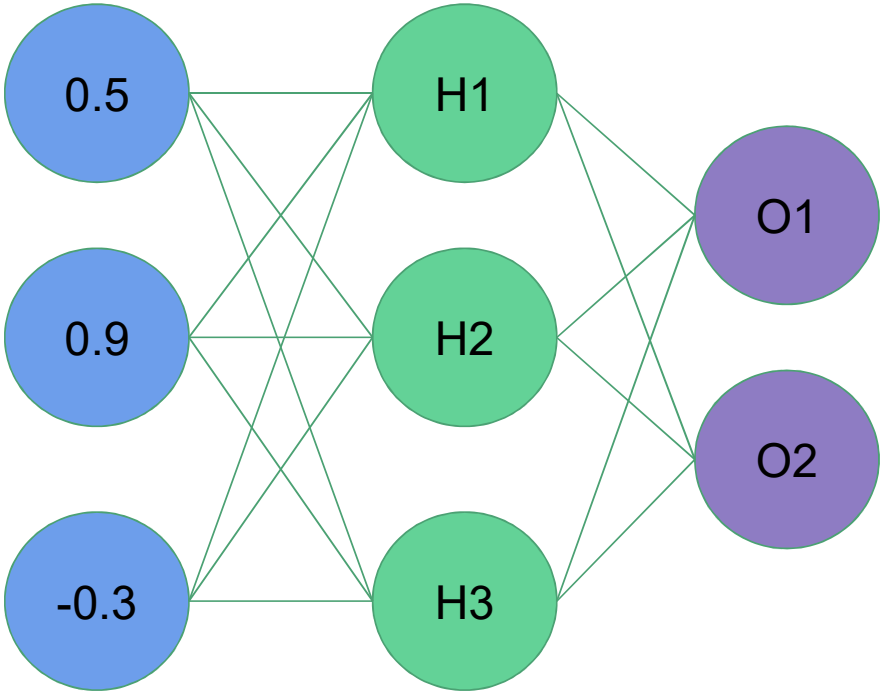
The California Stands By on Chance of Picking Up Other Boats or Rafts.

OLYMPIC SENDS THE NEWS

Only Ship to Flash Wireless Messages to Shore After the Disaster.

The Lost Titanic Being Towed Out of Belfast Harbor.

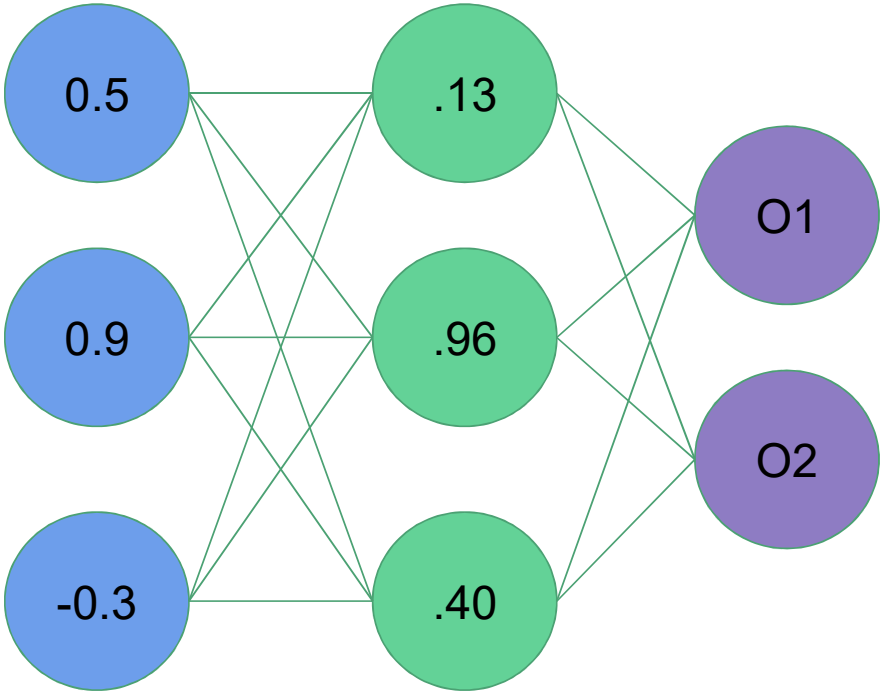
Inferenz
Neuronale Netze



H1 Weights = (1.0, -2.0, 2.0)
H2 Weights = (2.0, 1.0, -4.0)
H3 Weights = (1.0, -1.0, 0.0)

O1 Weights = (-3.0, 1.0, -3.0)
O2 Weights = (0.0, 1.0, 2.0)

Inferenz
Neuronale Netze

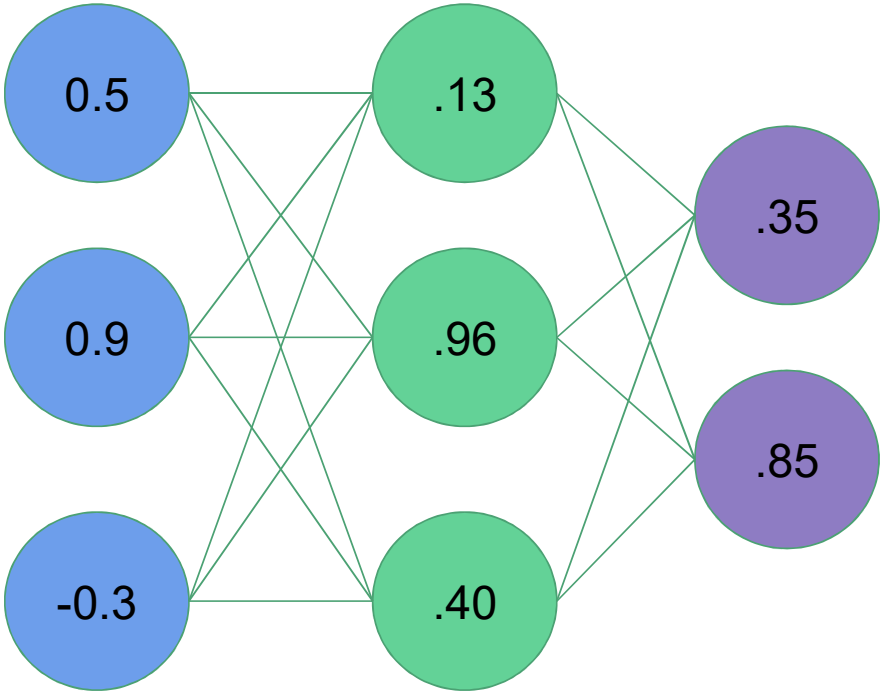


H1 Weights = (1.0, -2.0, 2.0)
H2 Weights = (2.0, 1.0, -4.0)
H3 Weights = (1.0, -1.0, 0.0)

O1 Weights = (-3.0, 1.0, -3.0)
O2 Weights = (0.0, 1.0, 2.0)

H1 = S(0.5 * 1.0 + 0.9 * -2.0 + -0.3 * 2.0) = S(-1.9) = .13
H2 = S(0.5 * 2.0 + 0.9 * 1.0 + -0.3 * -4.0) = S(3.1) = .96
H3 = S(0.5 * 1.0 + 0.9 * -1.0 + -0.3 * 0.0) = S(-0.4) = .40

Neuronale Netze
Inferenz



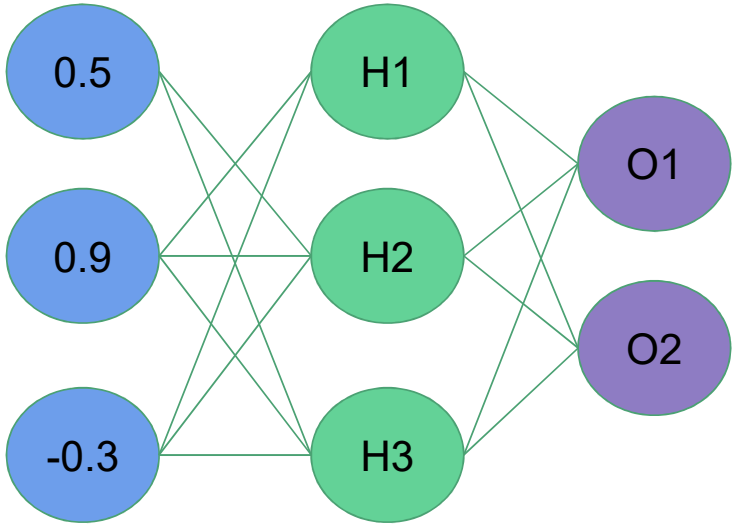
H1 Weights = (1.0, -2.0, 2.0)
H2 Weights = (2.0, 1.0, -4.0)
H3 Weights = (1.0, -1.0, 0.0)

O1 Weights = (-3.0, 1.0, -3.0)
O2 Weights = (0.0, 1.0, 2.0)

O1 = S(.13 * -3.0 + .96 * 1.0 + .40 * -3.0) = S(-.63) = .35
O1 = S(.13 * 0.0 + .96 * 1.0 + .40 * 2.0) = S(1.76) = .85

Matrix-Formulierung der Inferenz

Neuronale Netze



H1 Weights = (1.0, -2.0, 2.0)
H2 Weights = (2.0, 1.0, -4.0)
H3 Weights = (1.0, -1.0, 0.0)

Hidden Layer Weights | Inputs

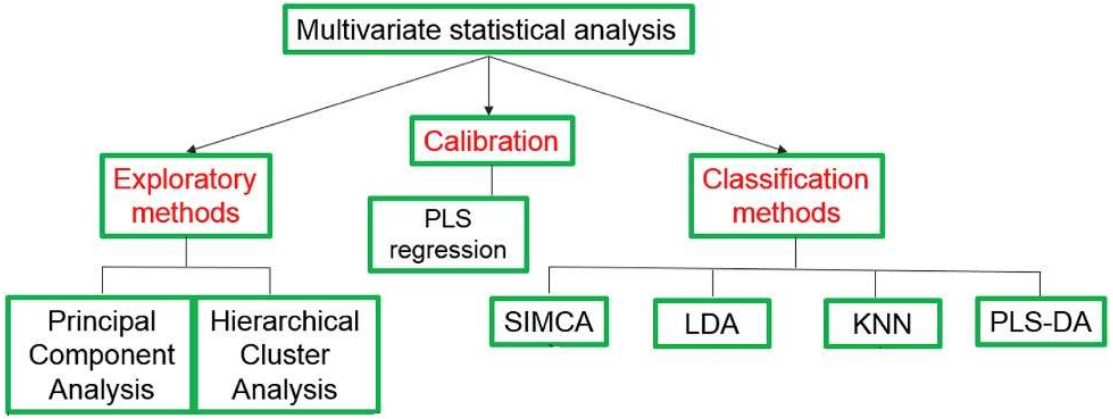
$$S\left(\begin{bmatrix} 1.0 & -2.0 & 2.0 \\ 2.0 & 1.0 & -4.0 \\ 1.0 & -1.0 & 0.0 \end{bmatrix} * \begin{bmatrix} 0.5 \\ 0.9 \\ -0.3 \end{bmatrix}\right) = S\left(\begin{bmatrix} -1.9 & 3.1 & -0.4 \end{bmatrix}\right) = \begin{bmatrix} .13 & .96 & 0.4 \end{bmatrix}$$

Hidden Layer Outputs

Unüberwachtes Lernen

„unsupervised learning“

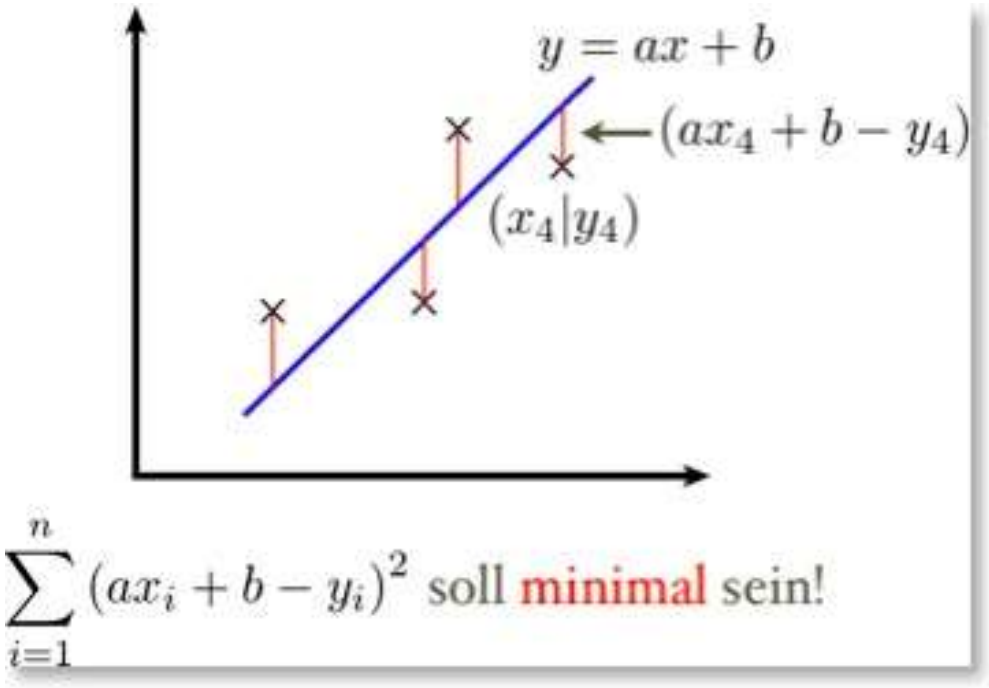
- Klassische Ballungen
  - k-means-Clustering
  - Agglomerative Hierarchical Clustering
- Begriffliche Ballungen („Conceptual Clustering“)
  - CLUSTER/2
  - Bildung von Begriffshierarchien („Concept Formation“)
- COBWEB
- CLASSIT
- Lernen durch Entdeckung
- BACON
- ABACUS
- ...
- Ohne Belohnung
- Ohne Vorwissen



Gauß'schen Methode der kleinsten Quadrate

Am Beispiel von 4 Messwerten

- Es soll eine passende Gerade (blau) gesucht werden.
- Nach einer Idee von **Carl-Friedrich Gauß** verwendet man als Maß für die Anpassung zunächst die vertikalen Abstände zwischen den tatsächlichen Messwerten und den entsprechenden Punkten auf der blauen Geraden.
- Für den vierten Punkt ist dieser Abstand angegeben.
- Diese Abstände werden quadriert und dann für alle Messwerte aufsummiert.
- Nach Gauß ist die beste Ausgleichslinie die, für die **die Summe der Quadrate der Abweichungen minimal ist**.
- Diese Idee heißt:
  - **Methode der kleinsten Fehlerquadrate**.
  - Das Stichwort „minimal“ deutet darauf hin, dass es sich um eine **Extremwertaufgabe** handelt.
  - Die Parameter a und b der Geraden müssen so gewählt werden, dass die bewusste Summe minimal wird.



Gütemaße für die Evaluation von Modellen

- Es existiert eine Reihe von Verfahren zum Testen der Güte statistischer Modelle.
- Die meisten solcher Verfahren liefern Kennzahlen, deren Quantität eine Aussage über die Modellgüte treffen. Ein Beispiel hierfür ist das Bestimmtheitsmaß R^2.
  - Dieses Maß erklärt den Anteil einer Variablen Y an der durch ein statistisches Modell erklärten Varianz.
  - Andere Gütetests beziehen sich auf die Beurteilung binärer Klassifikatoren.
  - Als Beispiel sei die Verifikation numerischer Wettervorhersagemodelle genannt.
  - An ihnen wird u.a. untersucht, wie genau ein vorhergesagter Parameter mit tatsächlich eingetretenen Ereignissen übereinstimmt (z.B. Niederschlag ja/nein).
  - Es werden Häufigkeitszahlen bestimmt, die angeben, wie oft eine Vorhersage bei eingetretenen Ereignissen korrekt war und wie oft inkorrekt.
  - Äquivalent dazu das Ganze für Nicht-Ereignisse.
  - Aus den erhaltenen vier Häufigkeiten lassen sich kategorische Gütemaße definieren.
  - Als Beispiele seien die Kennzahlen
    - Probability Of Detection
    - False Alarm Rate
    - True Skill Statisticsgenannt
  - Kontinuierliche Gütemaße beziehen sich dagegen auf kontinuierliche Parameter wie z.B. die Temperatur.
  - Hier findenübliche statistische Maße wiemittlerer Fehler,mittlerer absoluter Fehler,RMSE[1]unddieStandardabweichung Anwendung
  - Die Methodik, von der einige Verfahren erwähnt wurden, bezieht sich u.a. immer auf die Analyse eines kompletten Datensatzes, auf dem ein statistisches Modell beruht.
- Einen anderen Ansatzverfolgt dagegen die Kreuzvalidierung.
  - Ein Datensatz wird in mehrere Teile aufgeteilt.
  - Auf Basis eines Teils des Datensatzes wird ein statistisches Modellabgeleitet.
  - Das Modell wird auf den Rest des Datensatzes angewendet und die Abweichung zu den tatsächlichen Werten untersucht.
  - Das Verfahren dient also in erster Linieder Überprüfung derPrognosegüteeines Modells.

Gütemaße für die Evaluation von Modellen

Regression

- **R^2- Metric (R-Squared)**
  - Der Wert der R^2 kann Werte zwischen 0 und 1 annehmen
  - Nahe 1: Vorhersagen des Models sind gut
  - Nahe 0: Vorhersagen sind nicht besser als raten

- $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

$y_i = vorhergesagte\ Werte$

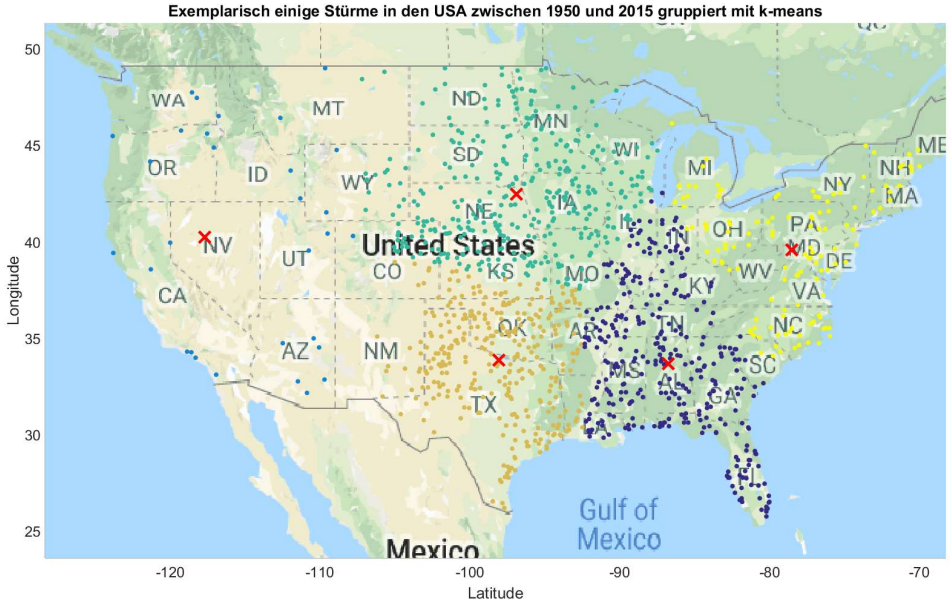
$\tilde{y}_i = korrekte\ Werte$

$\bar{y} = Median\ der\ Werte$

k-means

Partitionierend: Verfahren der Vektor-Quantisierung

- Bildung von Mengen ähnlicher Objekte in k Gruppen
  - mit möglichst geringe Varianz
  - In Gruppen ähnlicher Größen
- Datensatz in k Partitionen teilen, so dass Summe der quadrierten Abweichungen von Cluster-Schwerpunkten minimal wird



x_j: Datenpunkt
S_i: Cluster
μ_i: Schwerpunkt

$$J = \sum_{i=1}^{k} \sum_{x_j \epsilon S_i} \left\| x_j - \mu_i \right\|^2$$

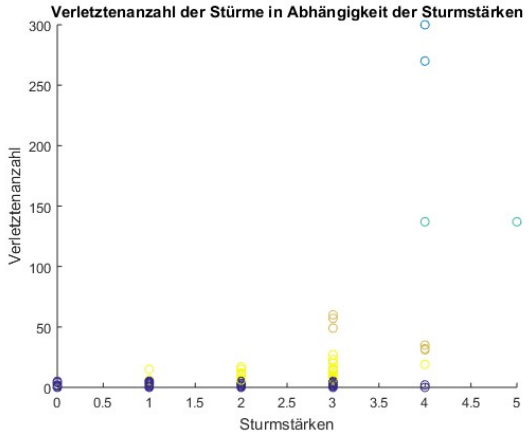
Lloyd Algorithmus:

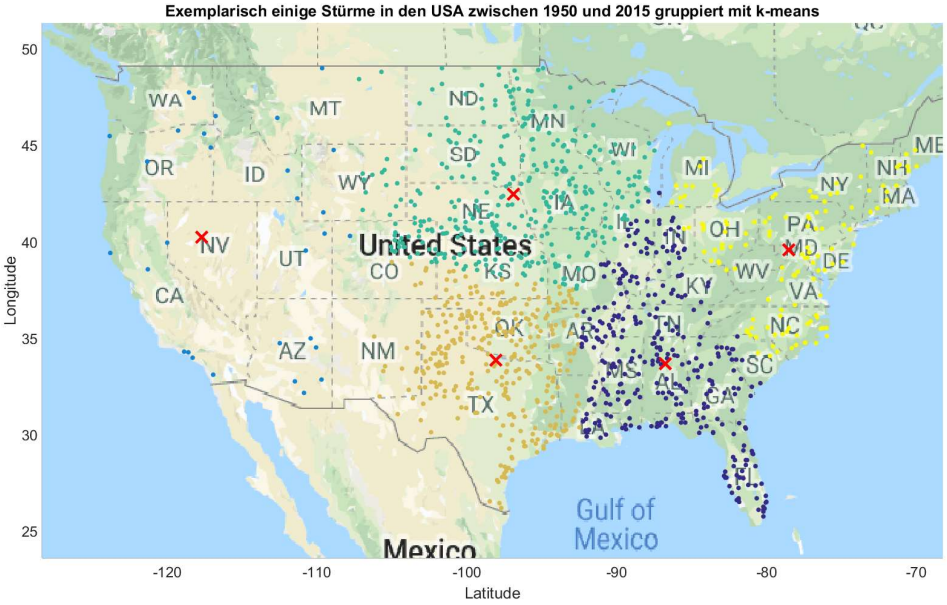
1. k unterschiedliche Zentren $c_1, c_2, \ldots, c_k$
2. Solange sich die Zielfunktion verbessert:
   Partitioniere P in Cluster $S_1, S_2, \ldots, S_k$ dass S_i die Punkte aus P enthält, deren nächstgelegenes Zentrum c_i ist
   Für jedes $1 \leq i \leq k$ sei $c_i \leftarrow \mu(C_i)$

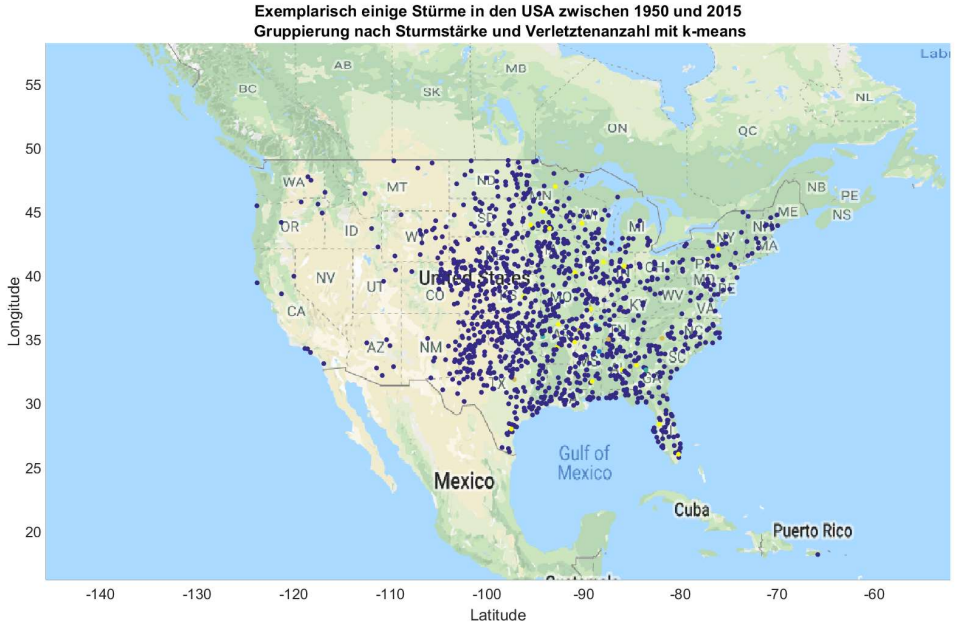
k-means

Partitionierend

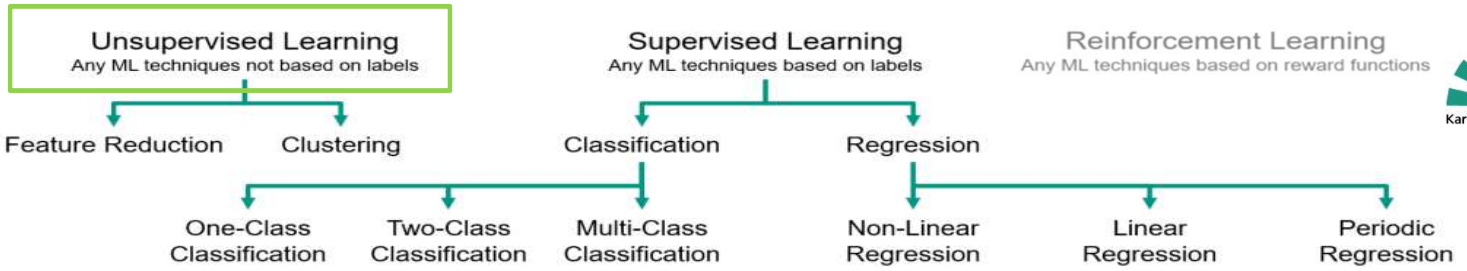
- Links:
  Cluster nach GPS Daten
- Rechts:
  Cluster nach Sturmstärken und Verletzten







Machinelles Lernen



Algorithm ↓	Abk.	Feature Reduction	Clustering	One-Class Classification	Two-Class Classification	Multi-Class Classification	Non-Linear Regression	Linear Regression	Periodic Regression
Faktorenanalyse		X							
Principal Component Analysis	PCA	X		X					
	K-means		X						
hierarchical cluster analysis	HCA		X						
DBSCAN			X						
One Class Support Vector Machine	OCSVM			X					
Isolation Forest				X					
	LODA			X					
(künstliche) Neuronale Netze	NN			X (Autoencoder)	X	X	X	X	X
Support Vector Machine	SVM				X				
Decision Tree					X	X			
Bayes-Klassifikation					X	X			
Random Forest						X			
Diskriminanzanalyse				x	x	x			
Logistic Regression							X	X	
Linear Regression								X	
Harmonic Regression									X
Nächste-Nachbar-Klassifikation	K-NN	x	x						

Nächste-Nachbar-Klassifikation

k-Nearest Neighbor

- Der nächste Nachbar (*Nearest Neighbor*) ist ein Maß der Entfernung multidimensionaler Datenpunkte mit kardinalen Variablenwerten.
- Ordinale und nominale Daten können transformiert werden, um so die Entfernung zu ermitteln.
  - Dieses Entfernungsmaß ist die Grundlage für die Nächste-Nachbar-Klassifikation (*k*-Nearest Neighbor oder *k*-NN), ein Verfahren, das vor allem als Klassifikationsverfahren Anwendung findet.
  - Es handelt sich um eine Vorgehensweise für die Mustererkennung in Daten, ohne dass „gelabelte“ Daten vorliegen.
  - Remark: Die Entfernung kann gewichtet oder ungewichtet berechnet werden.
- Ähnliche Fälle liegen nah beieinander → Fälle mit geringer Ähnlichkeit sind weit voneinander entfernt.
  - ► Fälle, die nahe beieinanderliegen, werden als Nachbarn bezeichnet.
  - ► Der Wert k bedeutet die Anzahl der Nachbarn.

Klassifizierung – Algorithmus

k-Nearest-Neighbor (KNN)

- Einfache Nearest-Neighbor Klassifikation (1NN)
  - Bestimmung nächster Nachbar über Distanz d

$$d(x, x_1) = \min\big(d(x, x_1)\big)$$

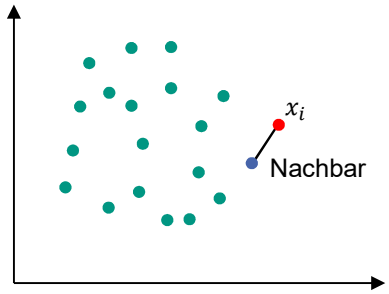
- Bestimmung der nächsten Nachbarn über

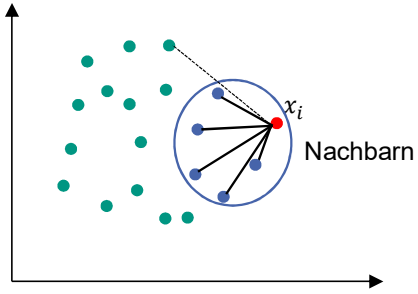
$$d(x, x_1) \leq d(x, x_2) \leq \cdots \leq d(x, x_n)$$

- x_i der Klasse zugeordnet, die unter k nächsten Nachbarn am häufigsten vorkommt
- Klassenwahrscheinlichkeit:

$$\hat{P}(y = g|x) = {}^1/_k \sum_{i=1}^{k} I(y_i = g)$$

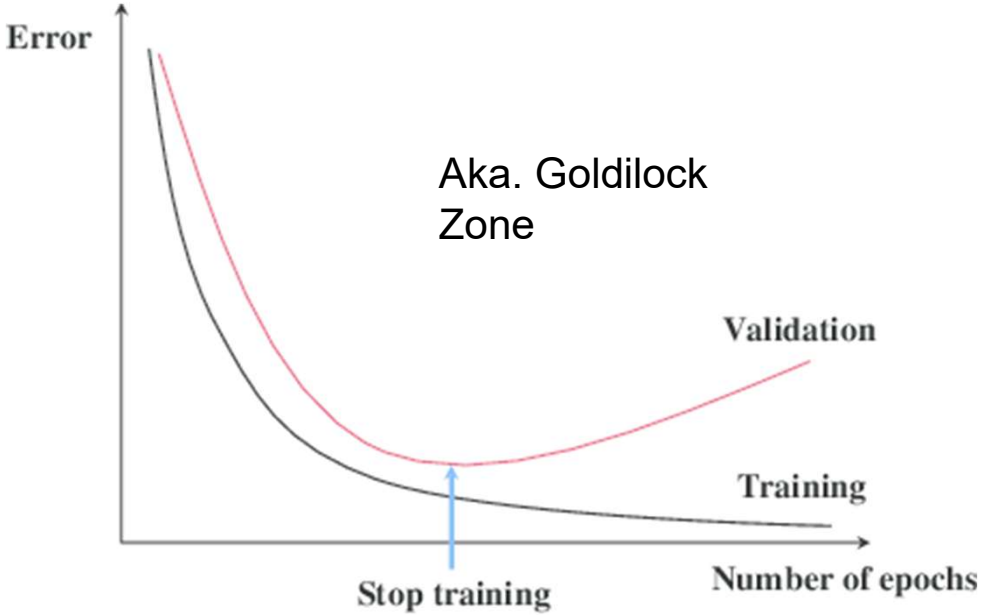
y: Klassenzugehörigkeit i-ter Nachbar
g: relative Häufigkeit einer Klasse





Regularisierung mit Early Stopping und Dropout

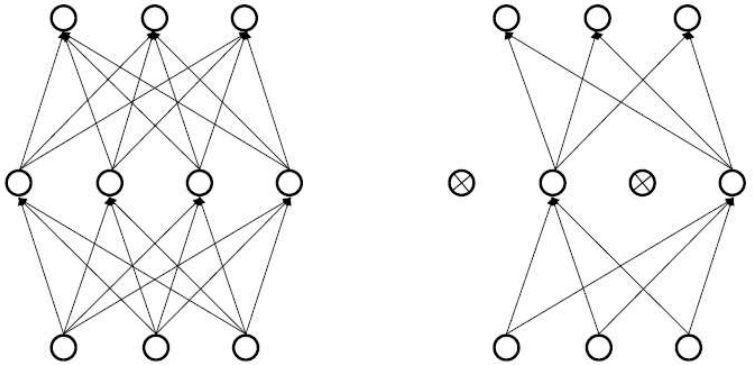
- Early Stopping
  - Grundidee ist es das Training abzubrechen, wenn sich die Genauigkeit nicht weiter verbessert



Aka. Goldilock Zone

- Dropout
  - Grundidee ist es die Komplexität eines KNNs zu reduzieren und Overfitting zu vermeiden
  - Entfernung einer bestimmten Anzahl von zufälligen Neuronen

Visualisierung von Dropout
(links: ohne Dropout, rechts: mit Dropout)



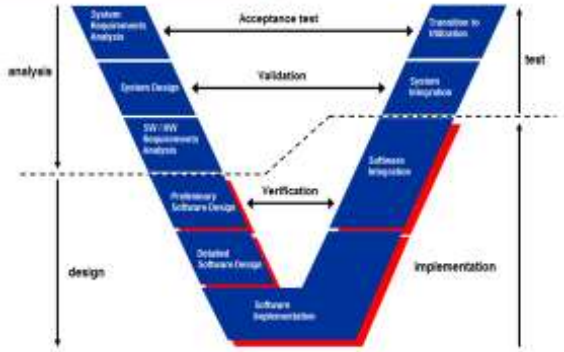
Evaluation

- Validierung entspricht der Prüfung der Business-Ziele
- Verifikation wird bei der Überprüfung des Prozesses durchgeführt

Verifikation und Validierung

Verifikation · Baue ich das Produkt richtig?

Validierung · Baue ich das richtige Produkt?



Business Understanding - Bestimmung der Data Mining Ziele

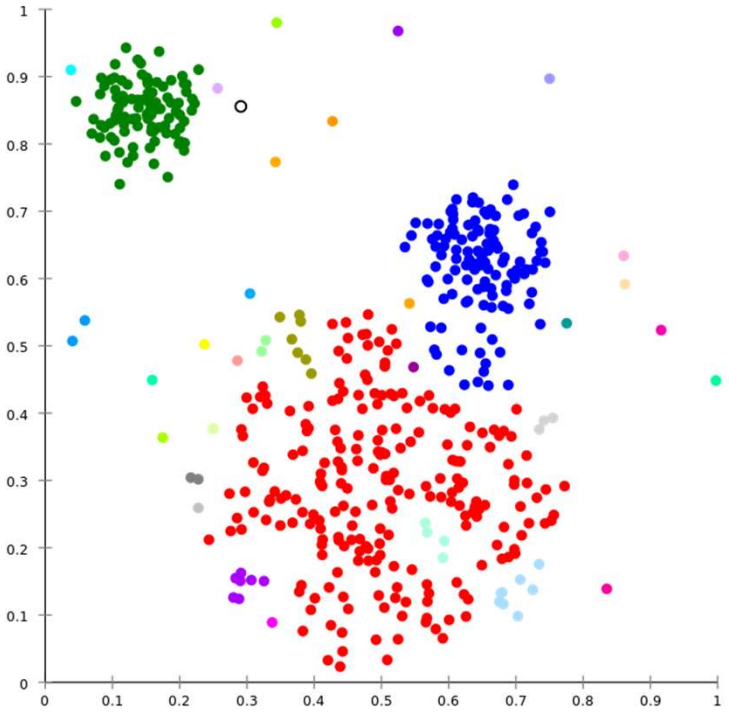
Während betriebswirtschaftlicher Erfolg im Sinne der wirtschaftlichen Terminologie Ziele beschreibt, definiert ein Data Mining Erfolg ein Ziel im technischen Sinne

- Data Mining Ziele
  „Übersetzung" der wirtschaftlichen Ziele in Data Mining Ziele
  Data Mining Ziel spezifizieren durch z.B. Klassifikation, Vorhersage, Beschreibung oder Ähnliches
- Data Mining Erfolgskriterien
  Definition der Kriterien für einen erfolgreichen Projektausgang im technischen Sinn wie z.B. Genauigkeit von Vorhersagen
  Kriterien zur Modellbewertung

Hierarchische Clusterverfahren

Agglomerative Clusteranalyse

- Für die Durchführung einer *agglomerativen* Clusteranalyse müssen …
  - … ein *Distanz-* oder *Ähnlichkeitsmaß* zur Bestimmung des Abstandes zwischen zwei Objekten
  - … ein *Fusionierungsalgorithmus* zur Bestimmung des Abstandes zwischen zwei Clustern ausgewählt werden.
- Wichtige Fusionierungsmethoden sind:
  - Single Linkage
    - Die Cluster, deren nächste Objekte die kleinste Distanz oder Unähnlichkeit haben, werden fusioniert.
  - Ward Methode
    - Die Cluster, die den kleinsten Zuwachs der totalen Varianz haben, werden fusioniert.



Single-linkage auf Normalverteilten Daten.
Durch den Single-link-Effekt trennen sich die beiden größten Cluster erst bei insgesamt 35 Clustern.