Einführung in Web- und Data-Science

Clustering

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

Danksagung

- Zur Vorbereitung dieser Präsentationen wurden Materialien verwendet von
  - Eamonn Keogh (University of California – Riverside) und
  - Sascha Szott (HPI Potsdam)

Clustering

- Form des unüberwachten Lernens
- Suche nach natürlichen Gruppierungen von Objekten
  - Klassen direkt aus Daten bestimmen
    - Hohe Intra-Klassen-Ähnlichkeit
    - Kleine Inter-Klassen-Ähnlichkeit
  - Ggs.: Klassifikation
- Distanzmaße
  - z. B. Minkowski Distanz (im $\mathbb{R}^n$):

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} = ||\mathbf{x} - \mathbf{y}||_p$$

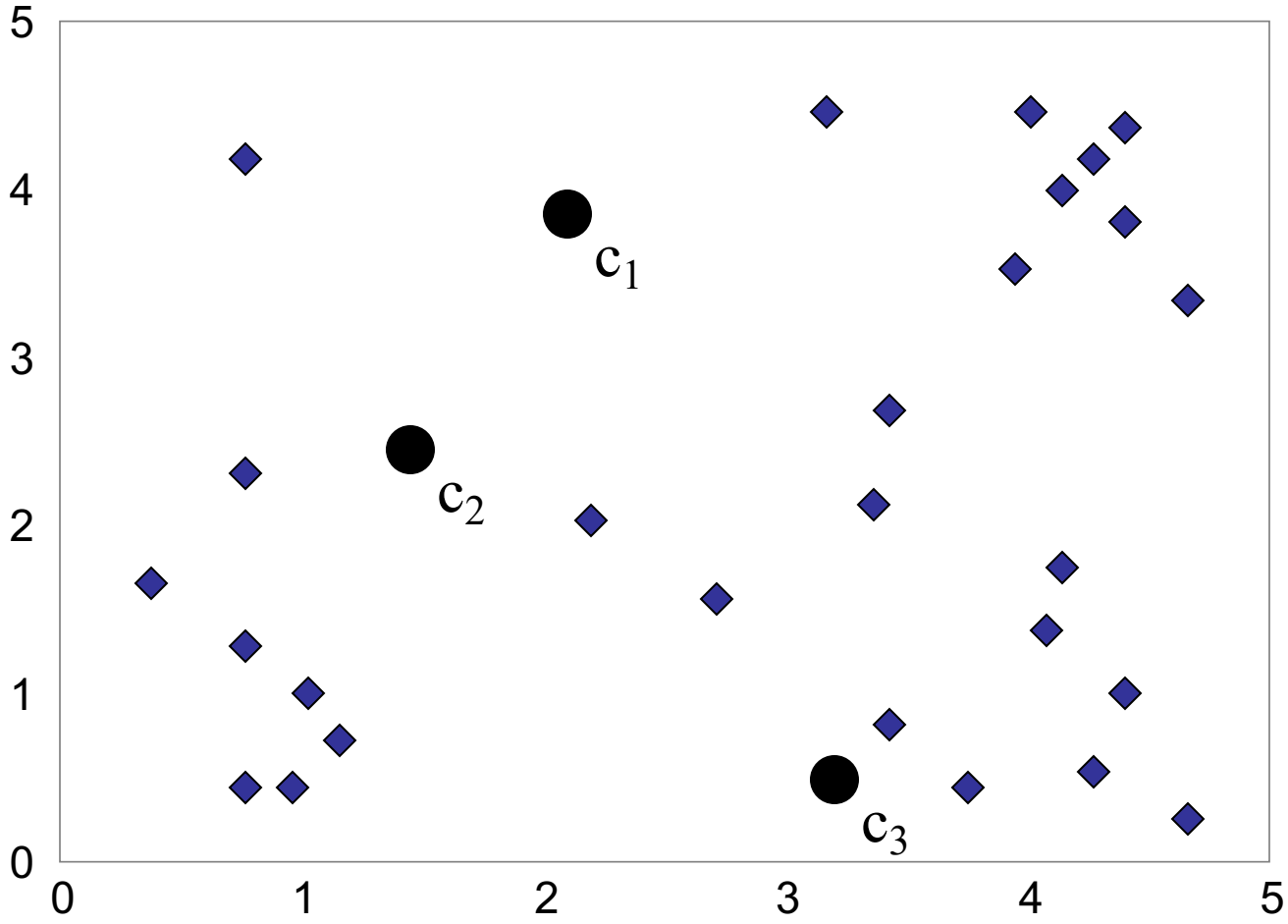
  - für $p = 1$: Manhattan Distanz
  - für $p = 2$: Euklidische Distanz

Clustering

K-MEANS

Partitionierung: K-means Clustering (1)

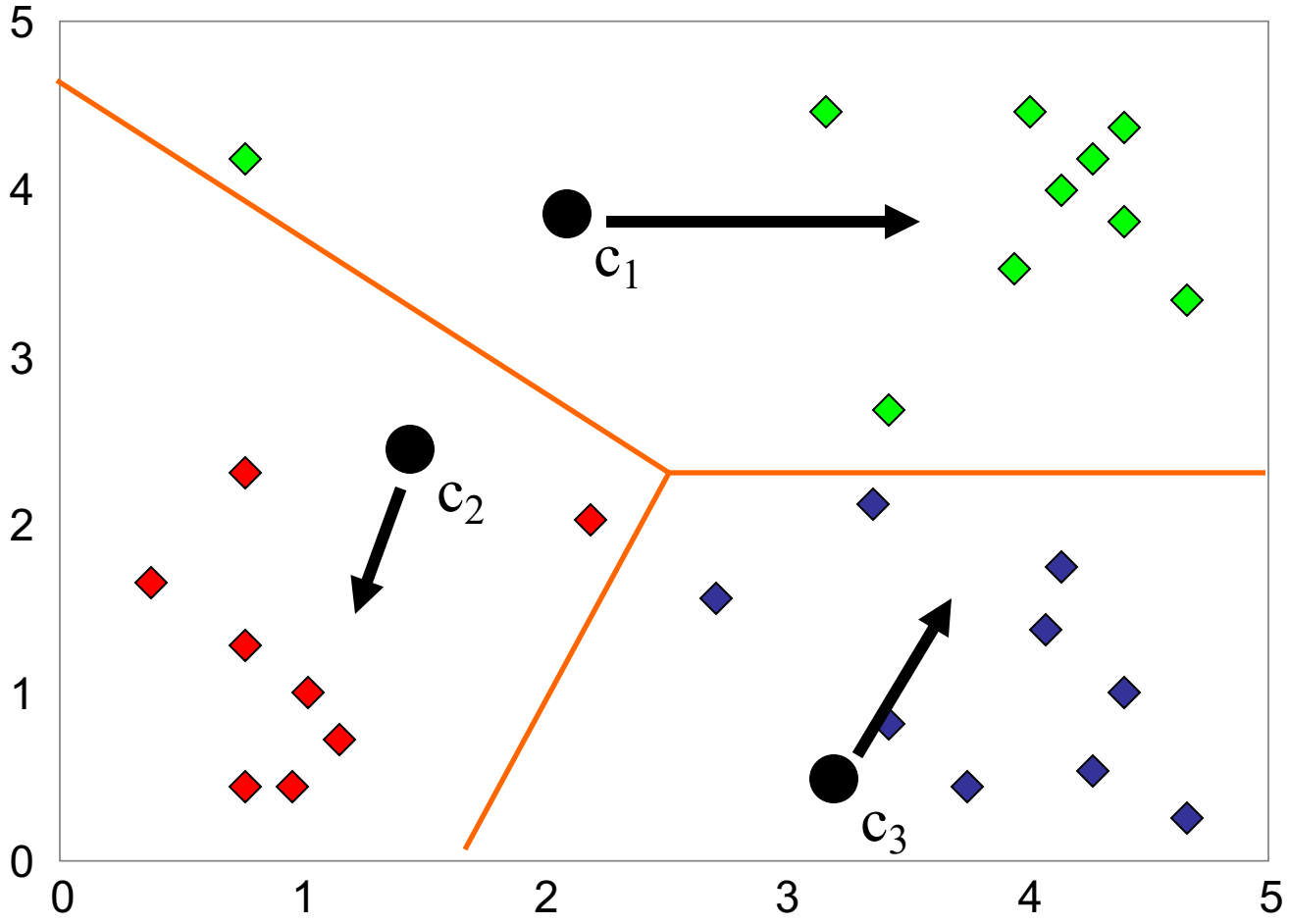
Distanzmaß: Euklidische Distanz



$$C_i^t = \left\{ x_j : \left\| x_j - c_i^t \right\|_2 \leq \left\| x_j - c_r^t \right\|_2 \text{ for all } r = 1 \ldots k, r \neq i \right\}$$

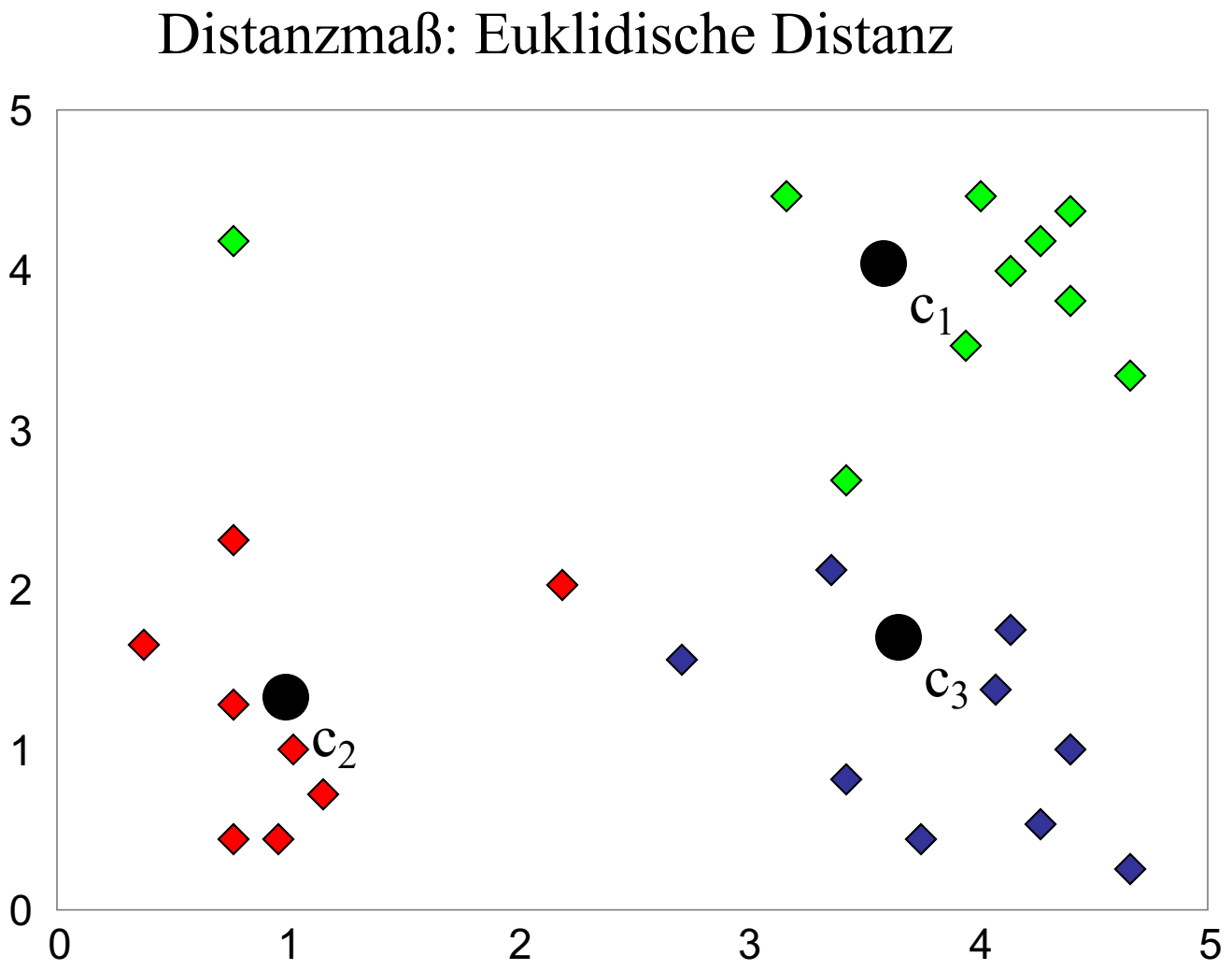
K-means Clustering (2)

Distanzmaß: Euklidische Distanz

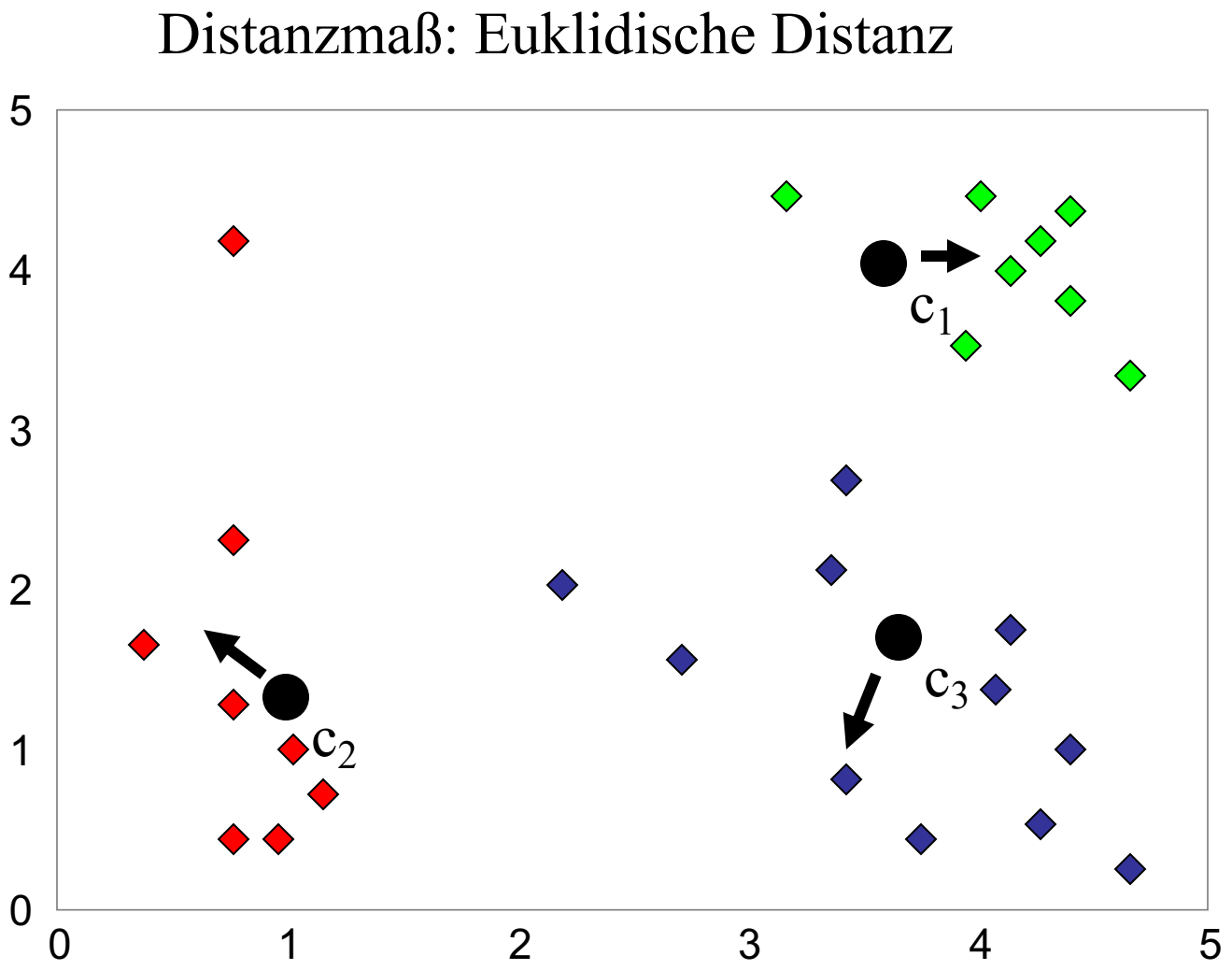


$$c_i^{t+1} = \frac{1}{|C_i^t|} \sum_{x_j \in C_i^t} x_j$$

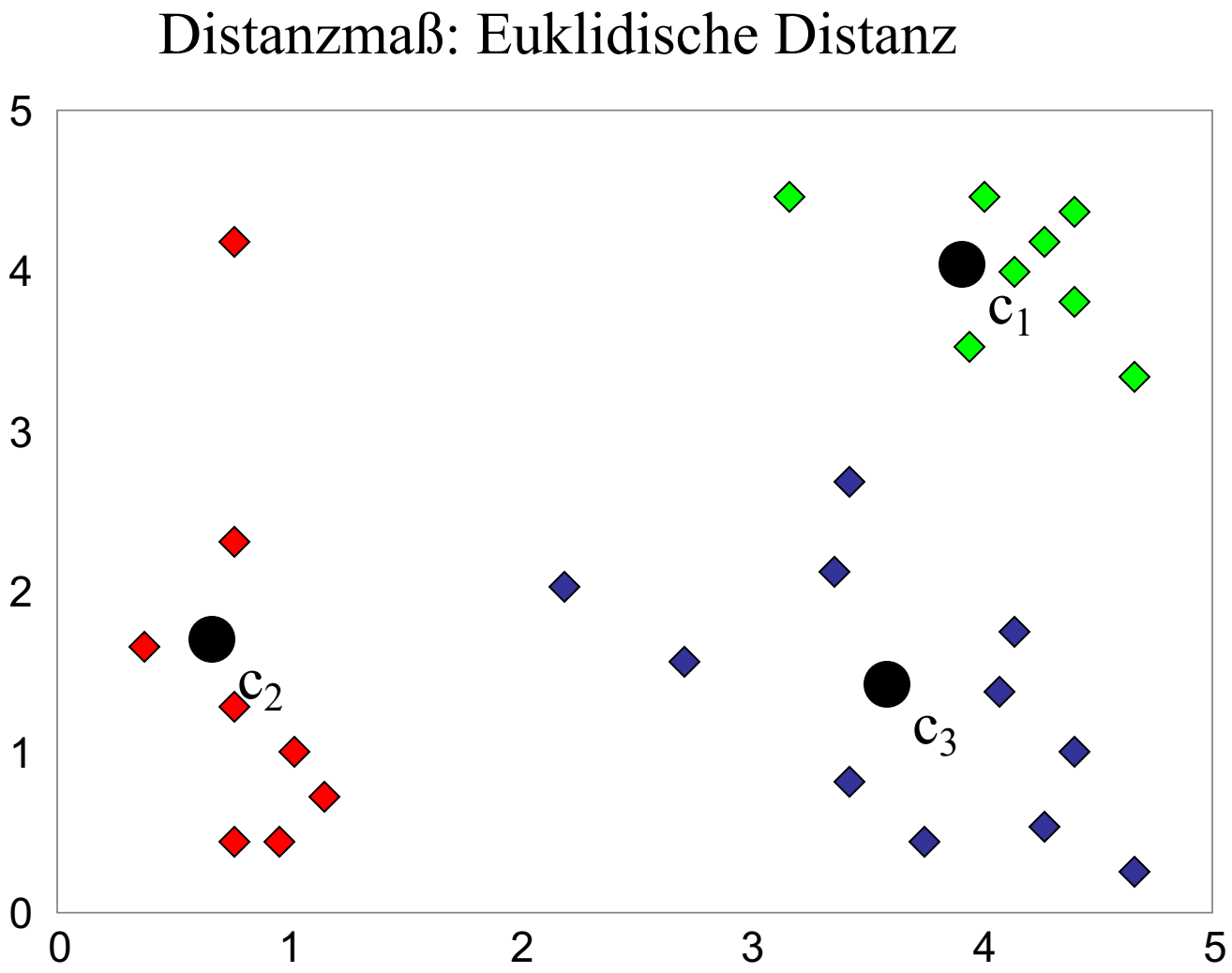
K-means Clustering (3)



K-means Clustering (4)



K-means Clustering (5)



K-Means: Cluster-Repräsentation

- Parameter $k \in \mathbb{N}$ bestimmt Anzahl der Cluster (woher?)
- Jedes Cluster C_i durch Zentroid $c_i \in \mathbb{R}^n$ repräsentiert Mittelwert bezüglich aller in C_i enthaltenen Punkte, d.h.,

$$c_i = \left(\frac{1}{|C_i|} \sum_{x_j \in C_i} x_j^1 , \dots , \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j^n \right)$$

  x_j^l l'te Kompontente

- Ziel: wähle Cluster $C_1, \dots, C_k \subseteq \mathcal{X}$ (alle Datenpunkte), so dass $\{C_1, \dots, C_k\}$ eine Partition von $\mathcal{X}$ ist und

$$E(C_1, \dots, C_k) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - c_i \right\|_2^2$$

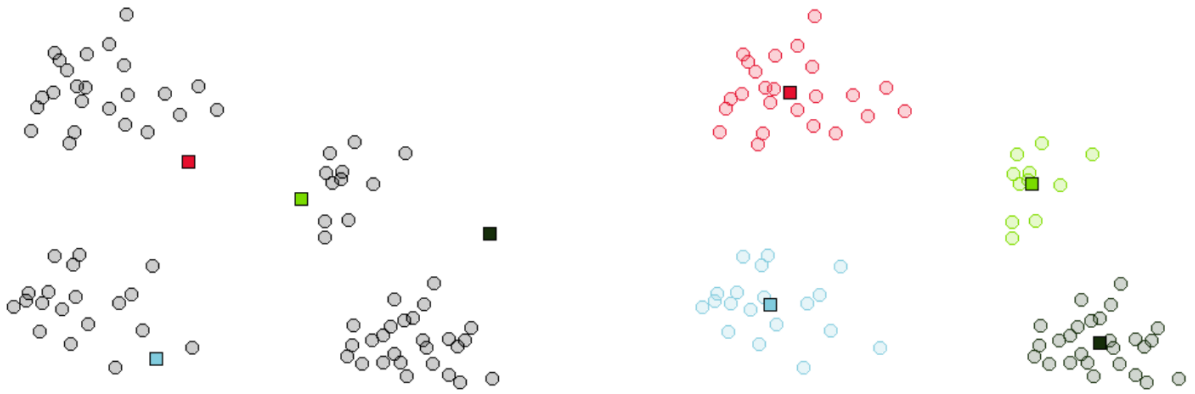
  (intra-cluster Varianz) minimiert wird

K-Means: Algorithmus

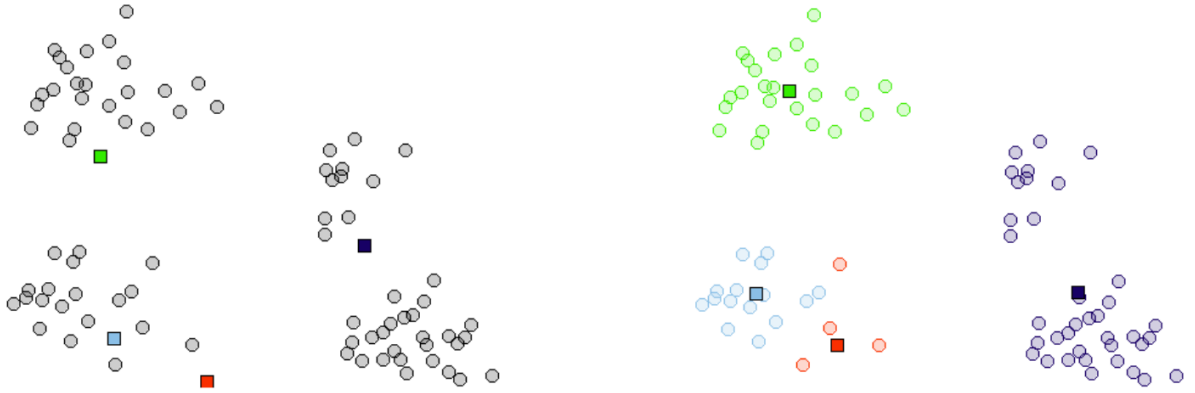
1. Wähle k zufällige Punkte $c_1, \dots, c_k \in \mathbb{R}^n$
2. $\forall\, x_j \in \mathcal{X}$: ordne x_j dem nähesten Zentroid zu, d.h., x_j wird c_i zugeordnet, falls
$$d(x_j, c_i) = \min_{1 \le i \le k} d(x_j, c_i)$$
wobei $d(.)$ eine Distanzfunktion ist (z.B. $\|.\|_2$)
3. Sei C_i die Menge aller Objekte, die c_i zugeordnet sind. Berechne ausgehend von C_i den Zentroid c_i neu.
4. Falls sich im vorherigen Schritt mindestens ein Zentroid geändert hat, gehe zu 2.
Andernfalls: Stop
   - $C_1, \dots, C_k$ ist eine Partitionierung von $\mathcal{X}$

K-Means-Ergebnis hängt vom Startwert ab

gutes Resultat:



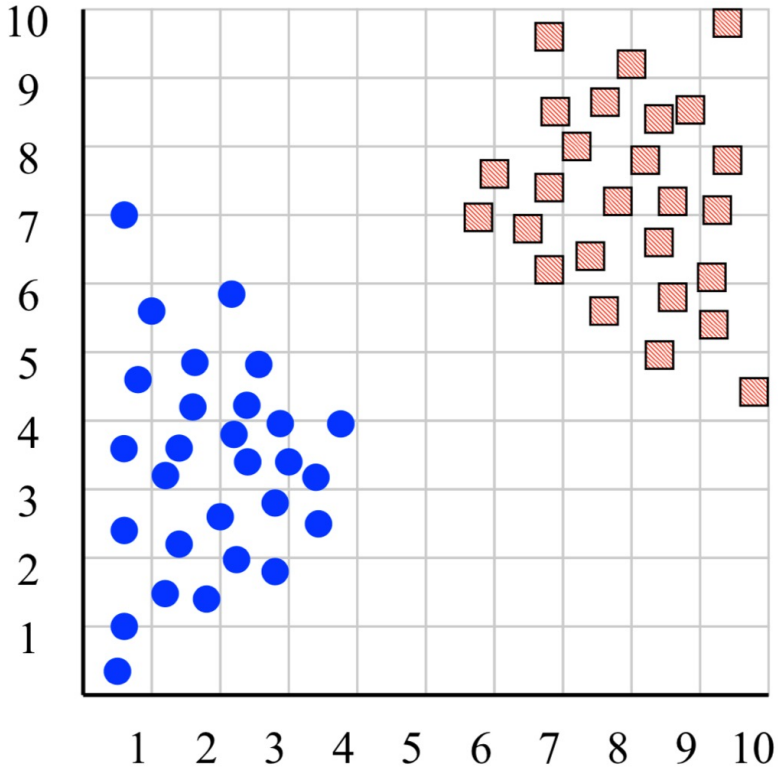
schlechtes Resultat:



Bestimmung von k

- Was ist die *richtige* Anzahl von Clustern?
- Schlecht gestelltes Problem
- Betrachten wir ein Beispiel (Grashüpfer/Heuschrecken-Datensatz)

Annahme:
Clusteranzahl nicht bekannt



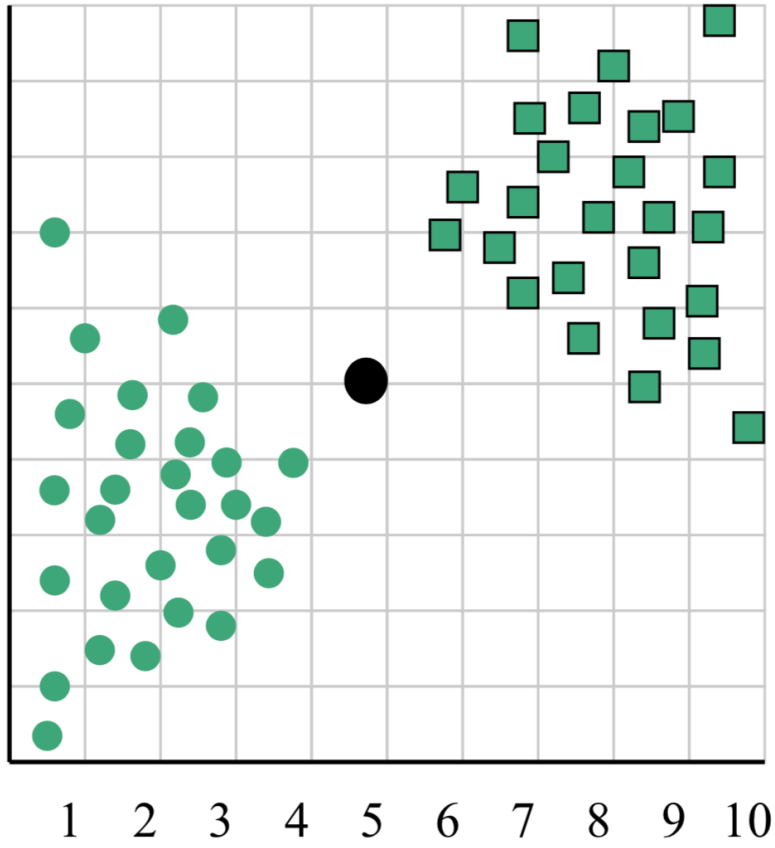
Bestimmung von k

Sei k = 1: $se_K =$ 873.0

$$se_{K_i} = \sum_{j=1}^{m_i} \| t_{ij} - C_i \|_2$$

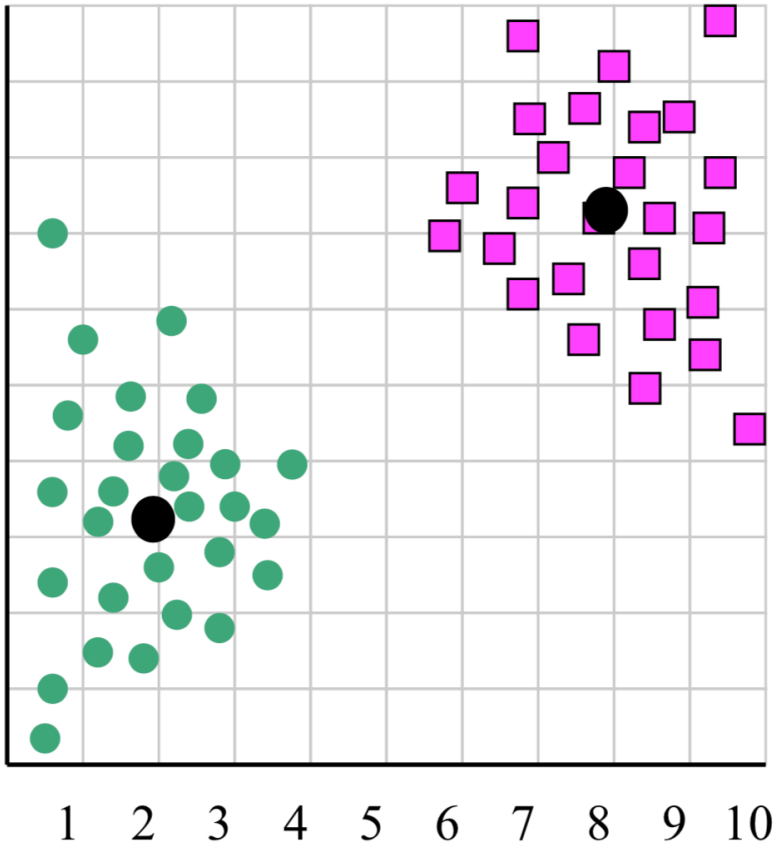
$$se_K = \sum_{j=1}^{k} se_{K_j}$$

se = squared error



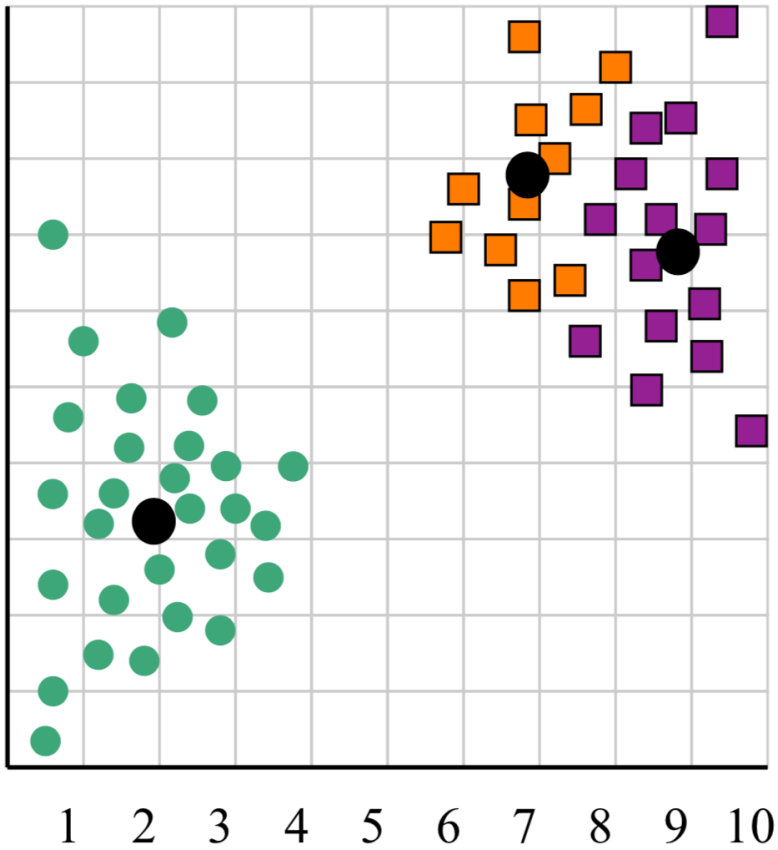
Bestimmung von k

Sei k = 2: $se_K =$ 173.1

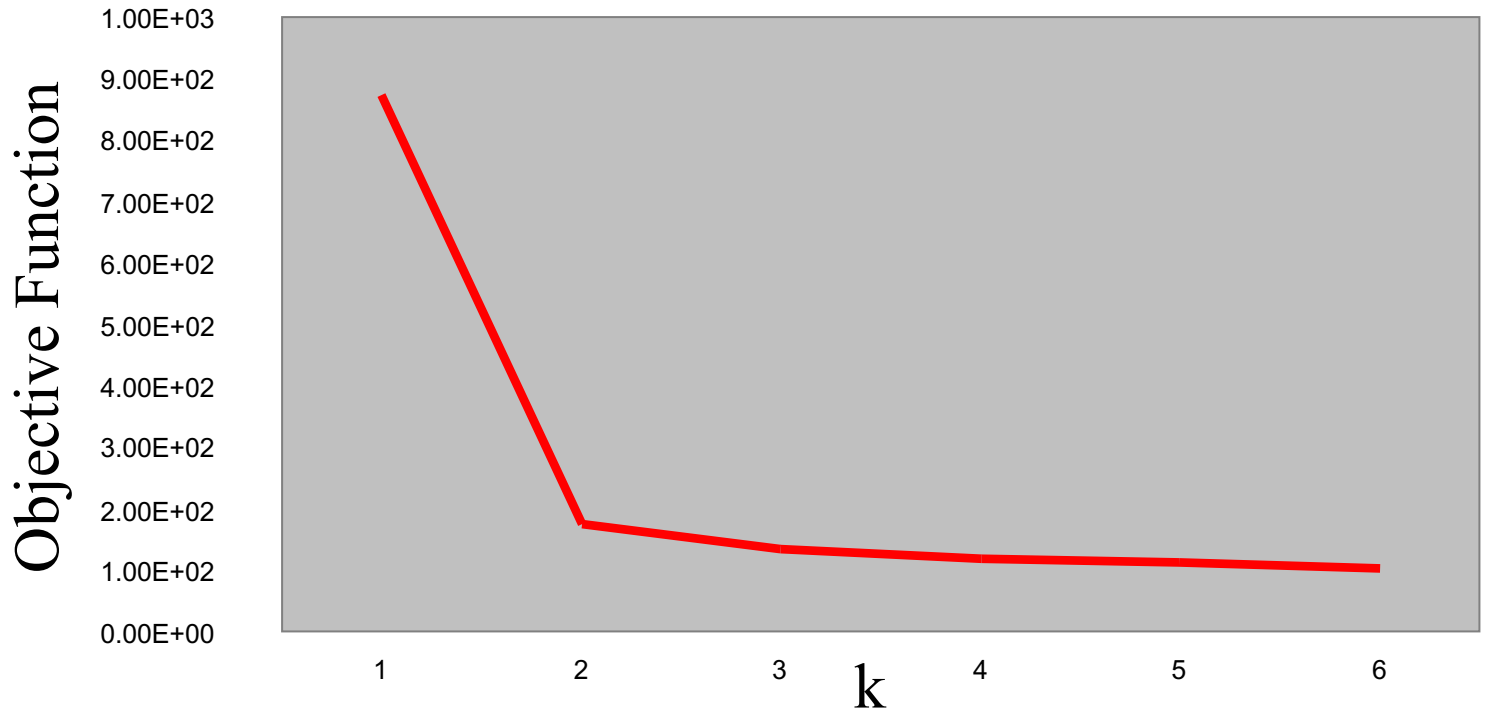


Bestimmung von k

Sei k = 2: $se_K =$ 133.6



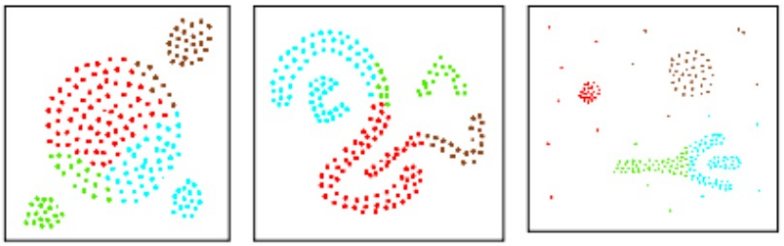
Was ist die richtige Clusteranzahl?



Variiere k und finde Knick in Graph der Bewertungsfunktion (Ellenbogen)

Diskussion

- Meist relativ wenige Schritte notwendig
  - Findet aber ggf. nur lokales Optimum
- Nur anwendbar, wenn Mittel definiert
  - Erweiterungen für kategoriale Daten existieren
- Basiert auf vorgegebener Clusteranzahl k
- Cluster haben meist gleiche Größe
- Probleme bei nichtkonvexen Formen
  - Varianten von K-Means (z.B. K-Medoid)



Trend

Wunsch

Clustering

DBSCAN

Dichtebasierendes partitionierendes Clustering

- DBSCAN-Verfahren (Density Based Spatial Clustering of Applications with Noise)
- Motivation: Punktdichte innerhalb eines Clusters höher als außerhalb des Clusters
- Resultierende Cluster können beliebige Form haben
  - Bei distanzbasierten Methoden ausschließlich konvexe Cluster
- Clusteranzahl k muss nicht initial vorgegeben werden

DBSCAN – Definitionen

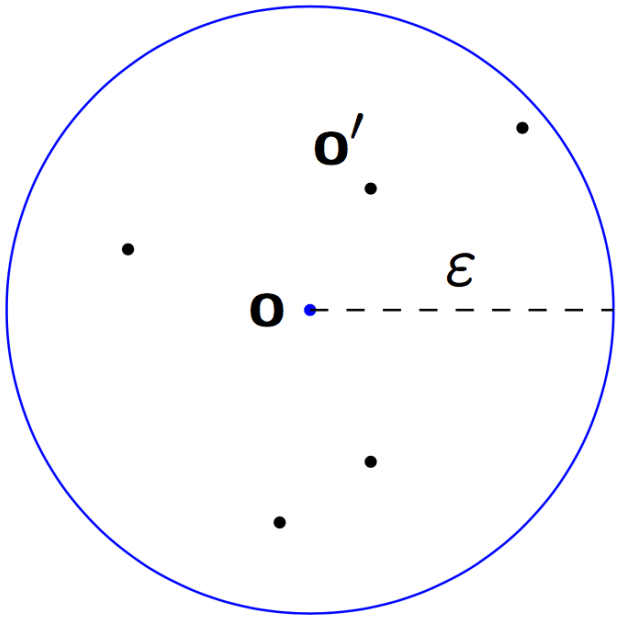
- ε-*Nachbarschaft* eines Objektes $\mathbf{o} \in O$:

$$N_\varepsilon(\mathbf{o}) := \{\mathbf{o}' \in O : d(\mathbf{o}, \mathbf{o}') \leq \varepsilon\}$$

- Aufteilung der Objekte in O
  - $\mathbf{o} \in O$ heißt *Kernobjekt* $:\Longleftrightarrow |N_\varepsilon(\mathbf{o})| \geq m$
  - $\mathbf{o} \in O$ heißt *Randobjekt* $:\Longleftrightarrow$ $\mathbf{o}$ ist kein Kernobjekt
- Parameter $\varepsilon \in \mathbb{R}^+$ und $m \in \mathbb{N}$ müssen initial vorgegeben werden (Heuristik zur Bestimmung der Parameter basierend auf der Dichte des „dünnsten“ Clusters)
- im Folgenden sei core(O) die Menge aller Kernobjekte in O
- in den folgenden Beispielen: $m = 4$

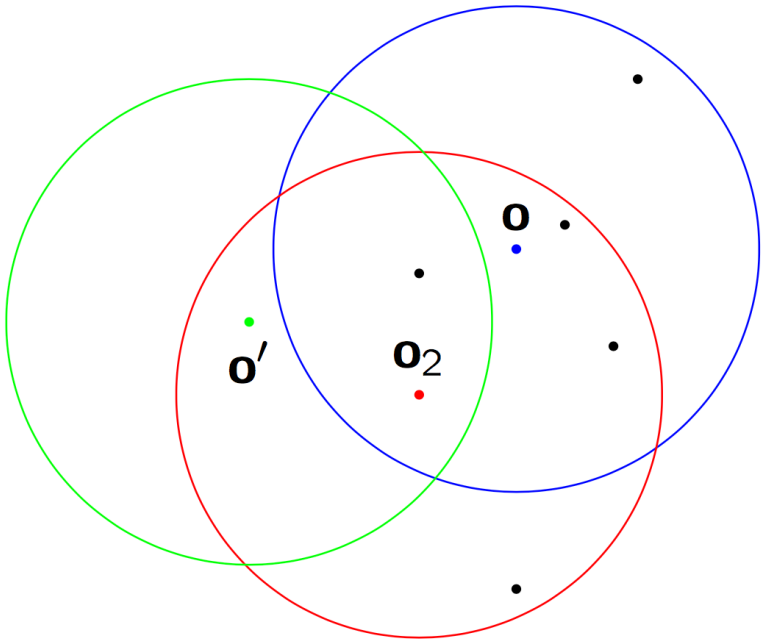
DBSCAN – Definitionen

$\mathbf{o}' \in O$ ist *direkt dichte-erreichbar* von $\mathbf{o} \in O :\Longleftrightarrow$
$\mathbf{o}' \in N_\varepsilon(\mathbf{o}) \wedge \mathbf{o} \in \text{core}(O)$



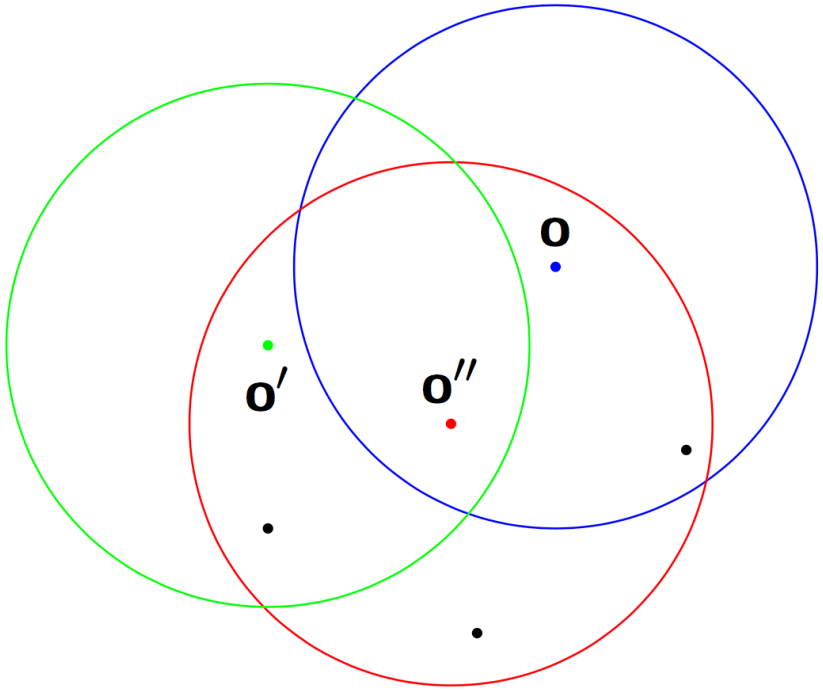
DBSCAN – Definitionen

$\mathbf{o}'$ ist *dichte-erreichbar* von $\mathbf{o}$ $:\Longleftrightarrow$
$\exists \mathbf{o}_1, \ldots, \mathbf{o}_i \in O : \mathbf{o}_1 = \mathbf{o} \wedge \mathbf{o}_i = \mathbf{o}' \wedge \forall j \in \{1, \ldots, i-1\} :$
$\mathbf{o}_{j+1}$ direkt dichte-erreichbar von $\mathbf{o}_j$



DBSCAN – Definitionen

$\mathbf{o}, \mathbf{o}' \in O$ sind *dichte-verbunden* $:\Longleftrightarrow \exists \mathbf{o}'' \in O$: $\mathbf{o}$ und $\mathbf{o}'$ sind von $\mathbf{o}''$ aus dichte-erreichbar



DBSCAN – Definitionen

Ein *Cluster* C ist eine nichtleere Teilmenge von O, die folgende Bedingungen erfüllt:

1. $\forall \mathbf{o}, \mathbf{o}' \in O$: ist $\mathbf{o} \in C$ und $\mathbf{o}'$ dichte-erreichbar von $\mathbf{o}$, dann ist $\mathbf{o}' \in C$ (Maximalität)
2. $\forall \mathbf{o}, \mathbf{o}' \in C$: $\mathbf{o}$ ist dichte-verbunden mit $\mathbf{o}'$ (Konnektivität)

Seien $C_1, \ldots, C_k$ Cluster bezüglich der Parameter (ε_i, m_i) mit $1 \leq i \leq k$. Dann ist die Menge N (*noise*) definiert als:

$$N := \{\mathbf{o} \in O : \forall i \in \{1, \ldots, k\}(\mathbf{o} \notin C_i)\}$$

N enthält also die Punkte, die keinem Cluster zugeordnet sind.

DBSCAN – Lemma 1

Lemma 1

Sei $\mathbf{o} \in \text{core}(O)$, dann ist die Menge
$\{\mathbf{o}' \in O : \mathbf{o}' \text{ ist dichte-erreichbar von } \mathbf{o}\}$ ein Cluster.

Bestimmung eines Clusters C in zwei Schritten

1. wähle einen beliebigen Punkt $\mathbf{o} \in \text{core}(O)$
2. ermittle die Menge P aller Objekte, die von $\mathbf{o}$ aus dichte-erreichbar sind

Dann ist $C = P \cup \{\mathbf{o}\}$.

DBSCAN – Lemma 2

Lemma 2

Sei C ein Cluster und $\mathbf{o} \in C$ ein Kernobjekt. Dann gilt folgende Gleichung

$$C = \{\mathbf{o}' \in O : \mathbf{o}' \text{ ist dichte-erreichbar von } \mathbf{o}\}.$$

Damit folgt, dass ein Cluster durch *jedes* beliebige seiner Kernobjekte eindeutig bestimmt ist.

DBSCAN

Eingabe: O, ε, m
Ausgabe: Funktion $c : O \to \mathbb{N}$, die jedem Objekt eine Clusternummer zuordnet

$c\_id := 1$ // -1: unclassified, -2: noise
$\forall \mathbf{o} \in O : c(\mathbf{o}) := -1$
$\forall \mathbf{o} \in O$ do
  if $c(\mathbf{o}) = -1$ then
    if ExpandCluster$(O, \mathbf{o}, c_id, \varepsilon, m)$ then
      $c\_id := c\_id + 1$
    fi
  fi
od

ExpandCluster

Eingabe: $O, \mathbf{o} \in O, c\_id, \varepsilon, m$
Ausgabe: Wahrheitswert true oder false

$S := neighborhood(O, \mathbf{o}, \varepsilon)$
if $|S| < m$ then // $\mathbf{o}$ ist ein Randobjekt
  $c(\mathbf{o}) := -2$
  return false
else // $\mathbf{o}$ ist ein Kernobiekt
  // bestimme alle Objekte, die von $\mathbf{o}$ aus dichte-erreichbar sind
  $\forall \mathbf{o}' \in S : c(\mathbf{o}') := c\_id$
  $S := S - \{\mathbf{o}\}$
  while $S \neq \emptyset$ do
    $\mathbf{o}' := S.getElement()$
    $R := neighborhood(O, \mathbf{o}', \varepsilon)$
    if $|R| \geq m$ then
      $\forall \mathbf{o}'' \in R$ do
        if $c(\mathbf{o}'') \in \{-1, -2\}$ then
          if $c(\mathbf{o}'') = -1$ then
            $S := S \cup \{\mathbf{o}''\}$
          endif
          $c(\mathbf{o}'') := c\_id$
        fi
      od
    fi
    $S := S - \{\mathbf{o}'\}$
  od
  return true
fi

Wann ist eine Gruppierung gut?

- Ideen für Bewertungsmaß (objective function)
  - Hohe Intra-Klassen-Ähnlichkeit
  - Kleine Inter-Klassen-Ähnlichkeit
- Formalisierung
  - Intra-Cluster-Varianz kleiner als Inter-Cluster-Varianz

Clustering

ANOVA

Und wenn die Cluster schon gegeben sind?

Beispiel: 25 Patienten mit Blasen auf der Haut
Behandlung: Methode A, Methode B, Placebo
Messwerte: # der Tage bis zur Abheilung der Blasen

Daten aus Studie [und Mittelwerte]:

- A: 5, 6, 6, 7, 7, 8, 9, 10 [7.25]
- B: 7, 7, 8, 9, 9, 10, 10, 11 [8.875]
- P: 7, 9, 9, 10, 10, 10, 11, 12, 13 [10.11]

Können wir sagen, dass Methode A die beste ist?

Sind die Differenzen der Mittelwerte signifikant?

Variation ZWISCHEN Gruppen vs. Variation IN Gruppen (clusters)

Analysis of variation notwendig: ANOVA

Was macht ANOVA?

In der einfachen Form (es gibt viele Erweiterungen) testet ANOVA folgende Hypothese:

H_0: Die Mittelwerte sind gleich (unterscheiden sich nicht)

H_a: Nicht alle Mittelwerte sind gleich,
der Unterschied ist signifikant

- Sagt nichts darüber, welche sich unterscheiden
- Muss durch multiple Vergleiche später herausgefunden warden

Unterschiedshypothese

Die Ausgangssituation

Zwei Variablen:
1 Kategorisch (Typ, Gruppe), 1 Quantitativ (Wert)

Frage: Hängen die (Mittel der) quantitativen Variablen von der Gruppe (gegeben durch kategoriale Variable) ab, in der sich das Objekt befindet?

Wenn die kategoriale Variable nur 2 Werte hat:

- 2-Stichproben-t-Test

ANOVA ermöglicht 3 oder mehr Gruppen

Annahmen von ANOVA

- Jede Gruppe annähernd normalverteilt
  - Dies können Sie Überprüfe, indem Sie sich Histogramme ansehen oder Annahmen verwenden
  - Kann vernünftig mit einigen Nicht-Normalverteilungen umgehen, aber nur ohne größeren Diskrepanzen
  - Standardabweichungen jeder Gruppe ungefähr gleich
  - Faustregel: Verhältnis von größter zu kleinster Standardabweichung muss kleiner als 2:1 sein

Standardabweichungsüberprüfung

Variable	treatment	N	Mean	Median	StDev
days	A	8	7.250	7.000	1.669
	B	8	8.875	9.000	1.458
	P	9	10.111	10.000	1.764

Vergleich der größten und kleinsten Standardabweichungen:

- größte: 1.764
- kleinste: 1.458
- 1.458 x 2 = 2.916 > 1.764

Notation für ANOVA

- n = Anzahl der Objekte insgesamt
- I = Anzahl der Gruppen
- $\overline{x}$ = Mittelwert für den gesamten Datensatz

Group i hat

- n_i = Anzahl der Objekte in der Gruppe i
- x_{ij} = Wert für Objekte j in Gruppe i
- $\overline{x_i}$ = Mittelwert für die Gruppe i
- s_i = Standardabweichung für die Gruppe i

Funktionsweise von ANOVA (Übersicht)

ANOVA misst zwei Variationsquellen in den Daten und vergleicht ihre relativen Größen

- Variation ZWISCHEN Gruppen (MSG, Mean Square b. Groups)
  Betrachten Sie für jede Gruppe die Differenz zwischen ihrem Mittelwert und dem Gesamtmittelwert

$$N^{-1} \sum_{x_i} (\bar{x_i} - \bar{x})^2$$

N: Normalisierungswert
(korrigiert: Freiheitsgrade)

- Variation INNERHALB von Gruppen (MSE, Mean Squared Error)
  Für jeden Datenwert x_j der Gruppe i betrachten wir die Differenz zwischen diesem Wert und dem Mittelwert der Gruppe

$$M^{-1} \sum_{x_{ij}} \left(x_{ij} - \bar{x_i}\right)^2$$

M: Normalisierungswert
(korrigiert: Freiheitsgrade)

F Statistik

Die ANOVA F-Statistik ist ein Verhältnis der Zwischengruppenvariaton geteilt durch die Variation innerhalb der Gruppe:

$$F = \frac{Between}{Within} = \frac{MSG}{MSE}$$

Ein großes F ist ein Beweis *gegen* H_0, da es anzeigt, dass es mehr Unterschiede zwischen Gruppen als innerhalb von Gruppen gibt (daher unterscheiden sich die Mittelwerte zwischen mindestens zwei Gruppen).

H_0: Die Mittel aller Gruppen sind gleich.

H_0 in Bezug auf Cluster:
Cluster sind schlecht
(Zentroide sind gleich)

Ein kleineres Beispiel

Angenommen, wir haben drei Gruppen (#groups = l)

- Group 1: 5.3, 6.0, 6.7
- Group 2: 5.5, 6.2, 6.4, 5.7
- Group 3: 7.5, 7.2, 7.9

Wir erhalten folgende Statistiken:

SUMMARY				
Gruppen	*Anzahl*	*Summe*	*Durchschnitt*	*Varianz*
Group 1	3	18	6	0,49
Group 2	4	23,8	5,95	0,176667
Group 3	3	22,6	7,53333333	0,123333

ANOVA Ausgabe

ANOVA						
Source of Variation	*sum of squares*	*df*	*mean square*	*F*	*P-value*	*F crit*
Between Groups	5,127333333	2	2,56366667	10,21575	0,008394	4,737416
Within Groups	1,756666667	7	0,25095238			
Total	6,884	9				

1 weniger als die Anzahl der Gruppen: l-1

1 weniger als die Anzahl der Objekte

Anzahl der Datenwerte - Anzahl der Gruppen: n-l (entspricht df für jede Gruppe addiert)

Berechnen der ANOVA F-statistic

			WITHIN		**BETWEEN**	
			difference:		difference	
		group	data - group mean		group mean - overall mean	
data	group	mean	plain	squared	plain	squared
5,3	1	6,00	-0,70	0,490	-0,44	0,194
6,0	1	6,00	0,00	0,000	-0,44	0,194
6,7	1	6,00	0,70	0,490	-0,44	0,194
5,5	2	5,95	-0,45	0,203	-0,49	0,240
6,2	2	5,95	0,25	0,063	-0,49	0,240
6,4	2	5,95	0,45	0,203	-0,49	0,240
5,7	2	5,95	-0,25	0,063	-0,49	0,240
7,5	3	7,53	-0,03	0,001	1,09	1,188
7,2	3	7,53	-0,33	0,109	1,09	1,188
7,9	3	7,53	0,37	0,137	1,09	1,188
TOTAL				1,757		5,106
TOTAL/df				**0,25095714**		**2,55275**

1.757/7 5.106/2

overall mean: 6.44 F = 2.5528/0.25025 = 10.21575

Ist F also groß genug?

F ist
Mean Square Between Group / Mean Square Within Group

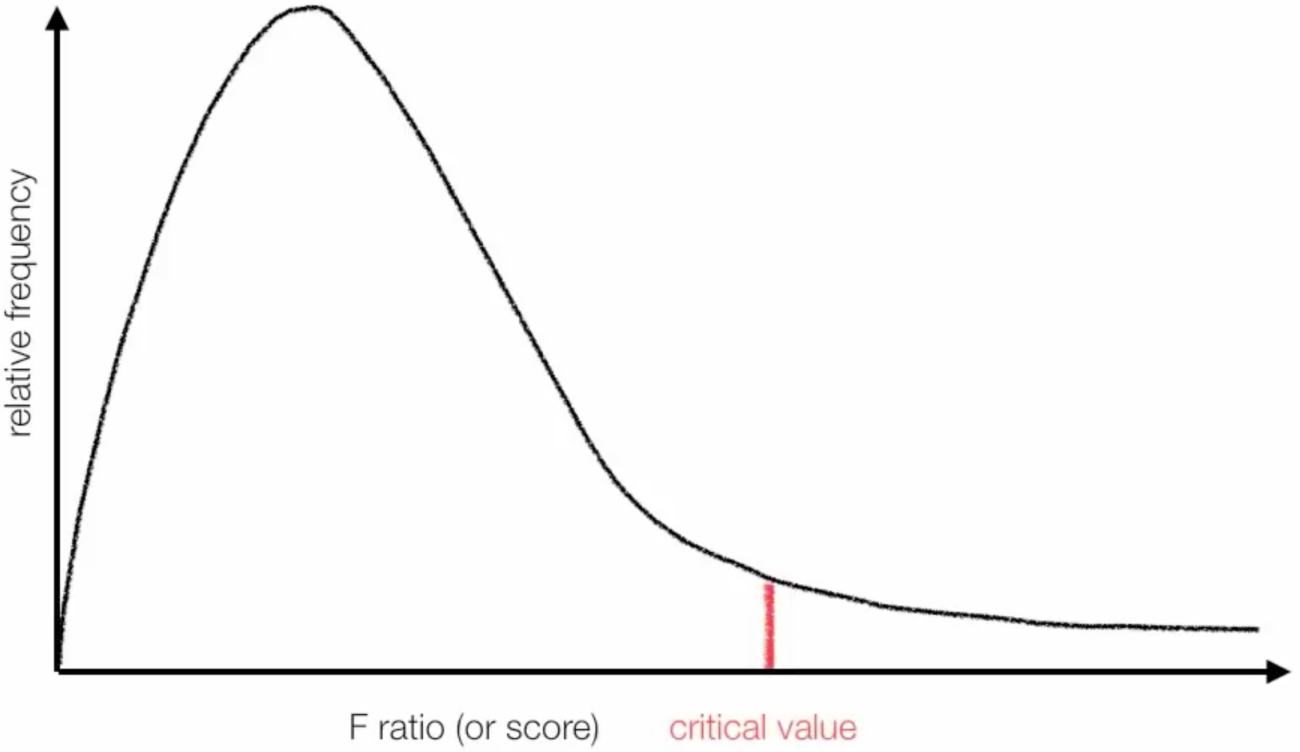
= MSG / MSE

Ein großer Wert von F bedeutet relativ mehr Unterschied zwischen Gruppen als innerhalb von Gruppen (Beweise gegen H0)

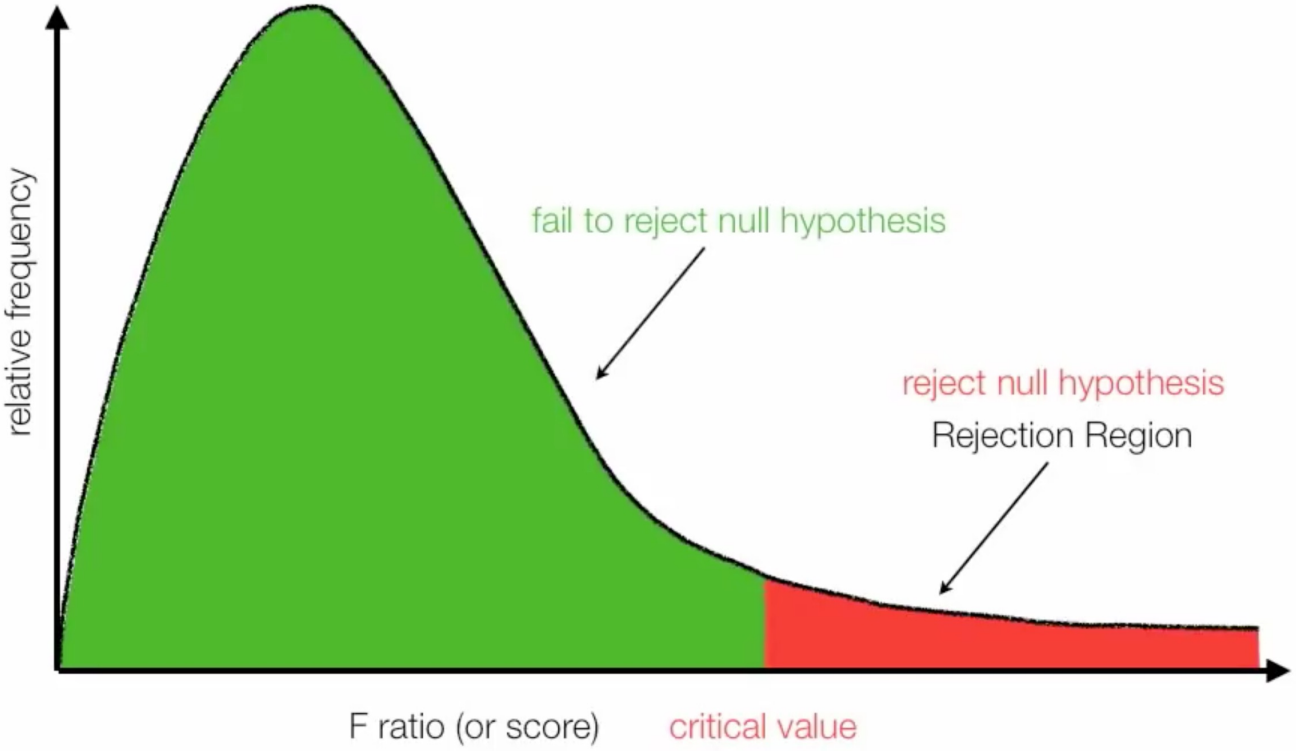
Um den p-Wert zu erhalten, vergleichen wir mit der *F(I-1,n-I)*-Verteilung

- *I-1* Freiheitsgrade im Zähler (# Gruppen -1)
- *n - I* Freiheitsgrade im Nenner (Rest der Freiheitsgrade)

F-Distribution



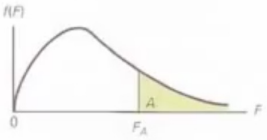
Kritischer Wert



F-Tabelle

$\alpha = 0.05$ (Verwenden Sie eine andere Tabelle für andere Werte von α)
Berechneter F-Value = 10.21
Kritischer Wert F(2, 7) = 4.74

Table 6(a) Critical Values of F: $A = .05$

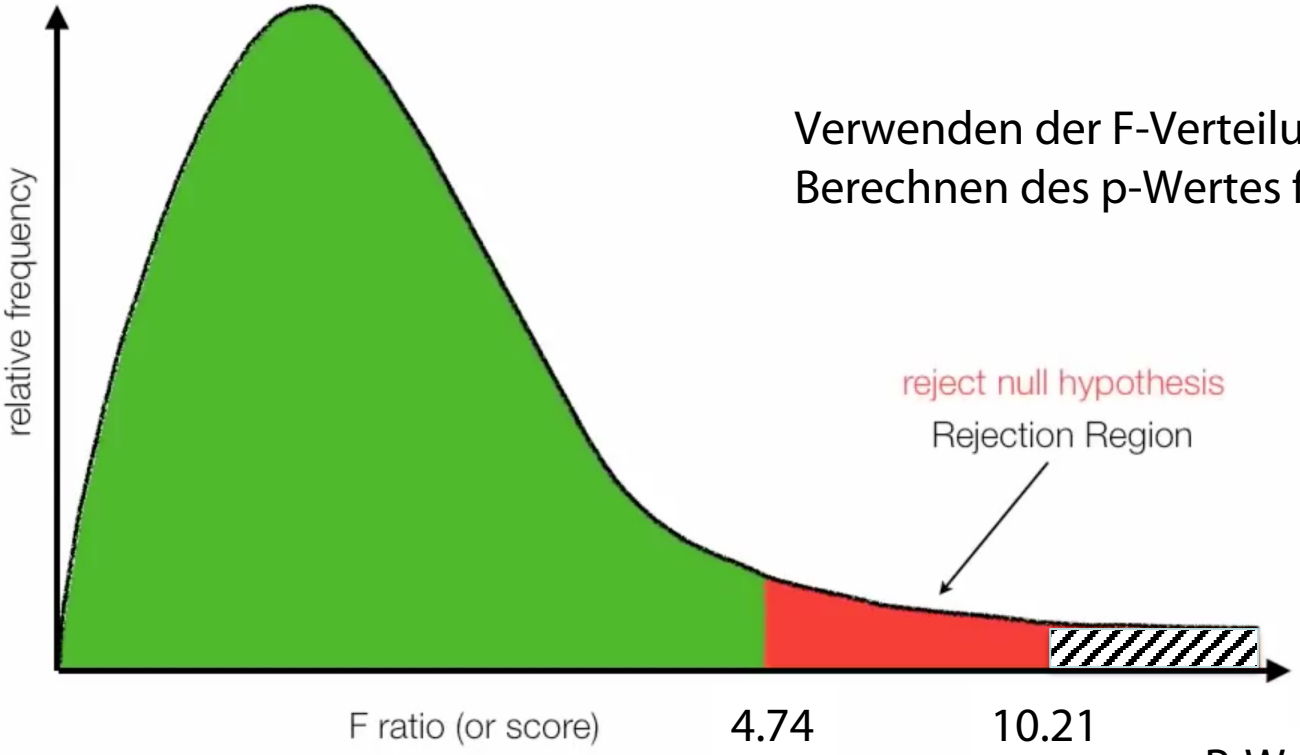


relates to groups or samples

relates to number of observations

ν_2 \ ν_1 (DENOMINATOR DEGREES OF FREEDOM / NUMERATOR DEGREES OF FREEDOM)	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28

Ablehnung der Nullhypothese



Verwenden der F-Verteilung zum Berechnen des p-Wertes für F=10.21

α=0.05 (rot+gestrichelte Fläche)

P-Wert=0.0084 (gestrichelte Fläche)

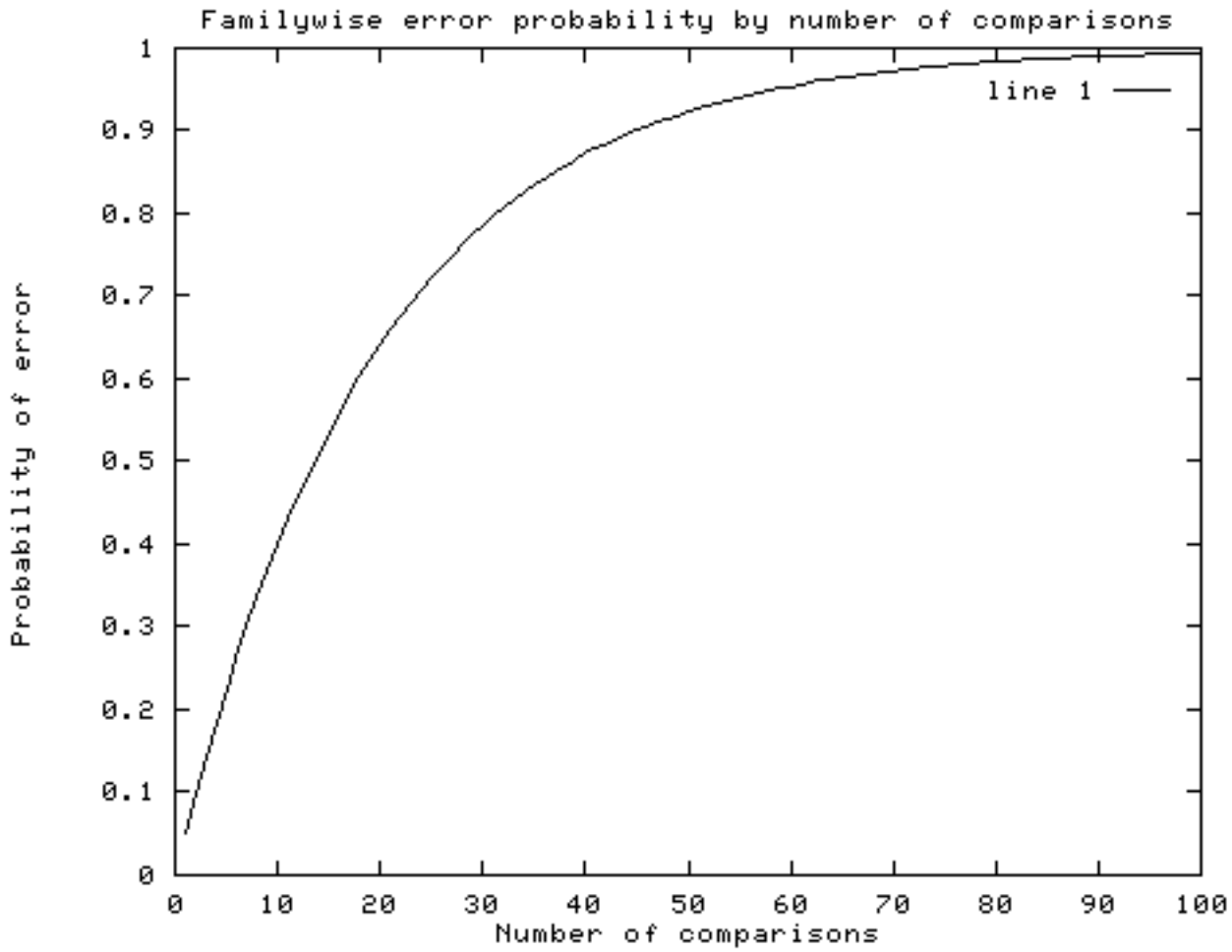
Warum nicht 3 paarweise t-Tests?

Antwort:

- Bei einer Fehlerquote von 5% für jeden Test (max Wert), ist die Gesamtwahrscheinlichkeit eines Typ-I-Fehlers bis zu $1-(.95)^3 = 14\%$
  - Wenn alle Vergleiche voneinander unabhängig sind
  - Für 6 Gruppen: ${}_6C_2 = 15$ paarweise T-Tests;
  - Hohe Chance, zufällig etwas signifikantes zu finden (wenn alle Vergleiche unabhängig wären mit einer Typ-I-Fehlerquote von jeweils 5 %)
  - Wahrscheinlichkeit von mindestens einem Typ-I-Fehler = $1-(.95)^{15}=54\%$.

$$ {}_nC_r = \binom{n}{k} = \frac{n!}{(n-k)!k!} $$

Mehrere Vergleiche



Korrektur für Mehrfachvergleiche

So korrigieren Sie mehrere Vergleiche post-hoc…

- Bonferroni-Korrektur (α um den konservativsten Betrag anpassen; Unter der Annahme, dass alle Tests unabhängig voneinander sind, dividieren Sie α durch die Anzahl der Tests)
- …

Bonferroni

Um beispielsweise eine Bonferroni-Korrektur vorzunehmen, dividieren Sie Ihren gewünschten Alpha-Cut-off-Wert (normalerweise 0,05) durch die Anzahl der Vergleiche, die Sie durchführen. Geht von völliger Unabhängigkeit zwischen Vergleichen aus, was viel zu konservativ ist.

Kritischer Wert für einseitigen Test	Ursprüngliches Alpha	# Tests	Neues Alpha
.001	.05	5	.010
.011	.05	4	.013
.019	.05	3	.017
.032	.05	2	.025
.048	.05	1	.050

Clustering

MANOVA

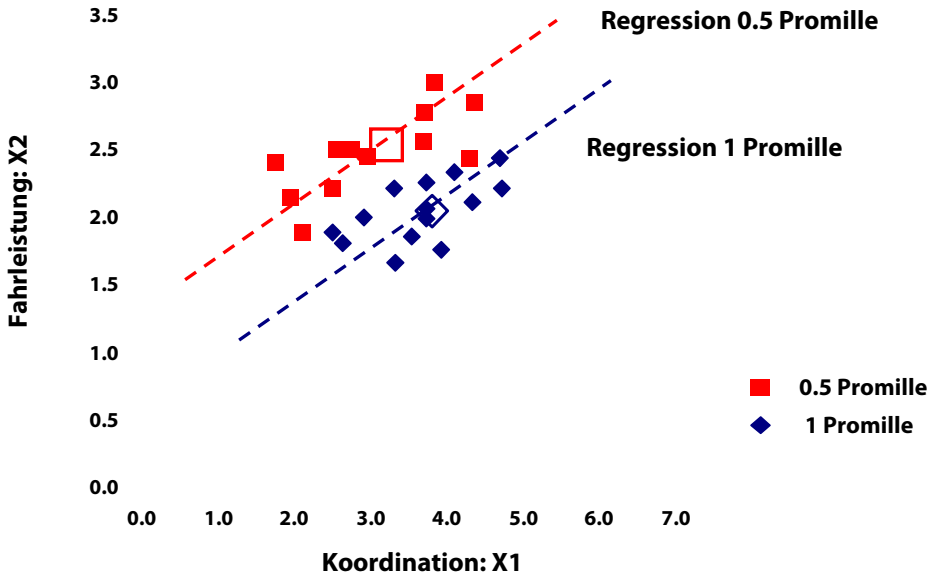
Multivariate Analysis of Variance: MANOVA

- Eine Erweiterung von ANOVA, in der die wichtigsten Effekte und Wechselwirkungen auf einer Kombination abhängiger Variablen bewertet werden
  - IV = Unabhängige Variable, manipulierte Variable (e.g., Behandlung)
  - DV = abhängige Variable, Messgröße (e.g., Mittel)

- MANOVA testet, ob mittlere Unterschiede zwischen Gruppen bei einer Kombination von DVs wahrscheinlich zufällig auftreten
- Es werden neue DVs erstellt, die lineare Kombinationen der einzelnen DVs sind, so dass der Unterschied zwischen den Gruppen maximiert wird
- Die Fragen sind meistens die gleichen wie bei ANOVA nur auf den linear kombinierten DVs statt nur einem DV

MANOVA

2D-Beispiel

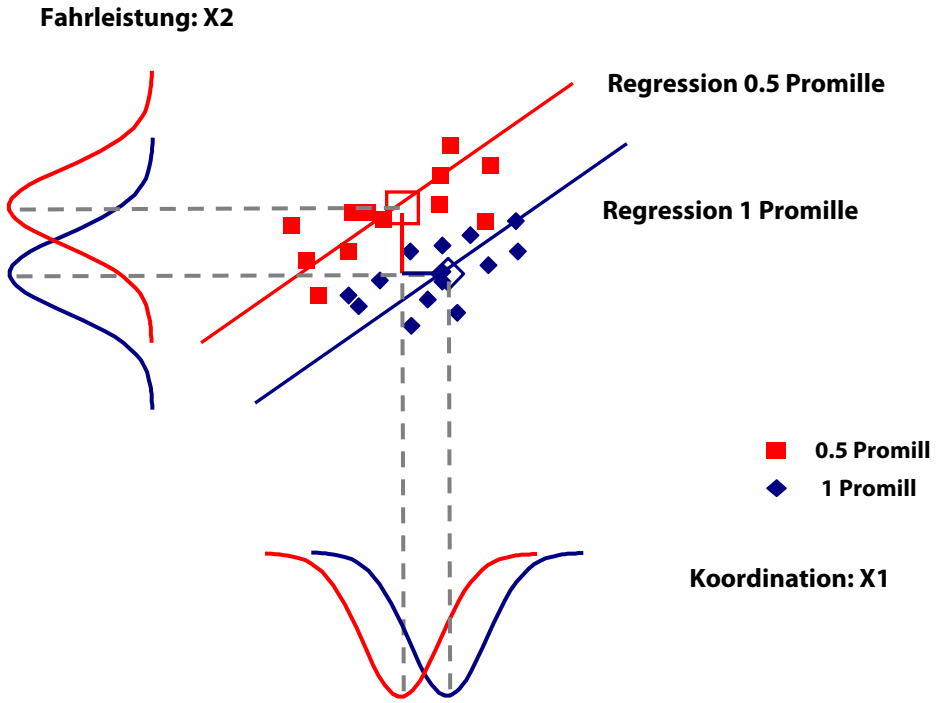


Prototypische Datensituation

- Generell: g- Gruppen gemessen auf p Variablen. Hier g=2, p=2, Koordination (X_1) und Fahrleistung (X_2)
- Gleiche Regressionssteigungen und gleiche Varianzen in den Gruppen auf beiden Variablen (Homogenität der Varianzen und Kovarianzen)
- Stichprobendaten entstammen multivariat normalverteilten Populationen.

MANOVA

2D-Beispiel

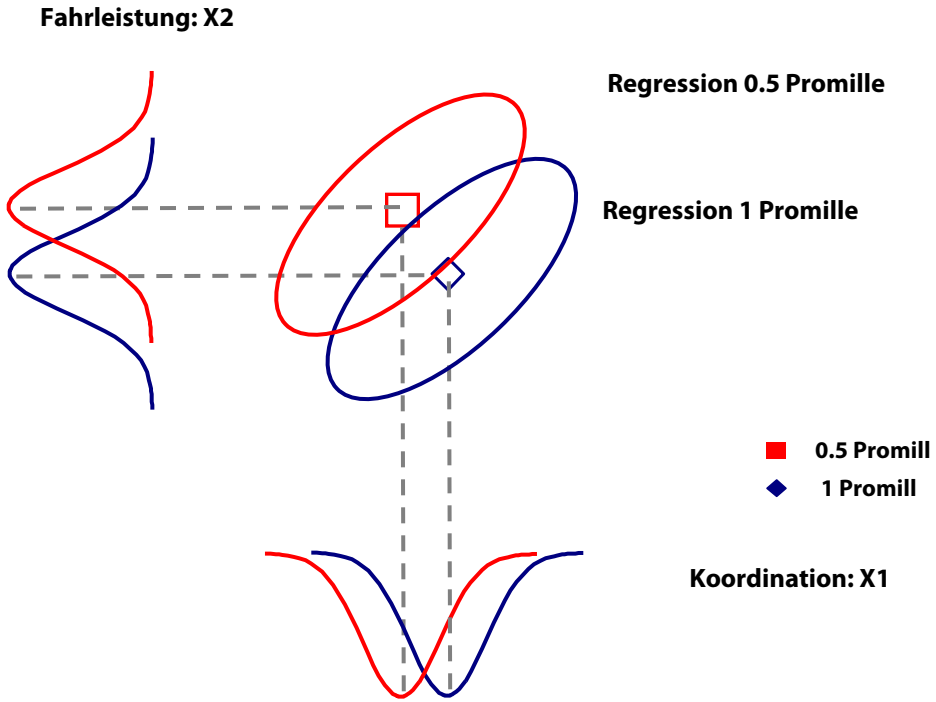


1D - Testen unzulänglich

- Univariat sind die Rohwertverteilungen nicht gut getrennt, und daher ebenfalls nicht die Mittelwerteverteilungen (hohes N nötig für signifikante Gruppenunterschiede in den Kennwerteverteilungen)
- Signifikanzurteile sind unabhängig und führen zu p Signifikanzaussagen, obwohl nur eine erwünscht ist
- Information der gleichen Beziehung zwischen den abhängigen Variablen (gleiche Korrelation) wird nicht genutzt .

MANOVA

2D-Beispiel



2D - Testen Ausgangslage

- 2D 95% Quantile zeigen an, dass die Mittelwerte der jeweils anderen Gruppe nicht mehr im Konfidenzbereich der Rohwerte liegen (bei den univariaten Verteilungen liegen sie darin)
- Orthogonal zur Hauptvarianzrichtung der Ellipsen bestehen optimale Trennbedingungen für die Mittelwerte
- Ein Test, in den die Korrelation der beiden Variablen eingeht, hat daher optimale Chancen, Unterschiede der Centroide aufzudecken.