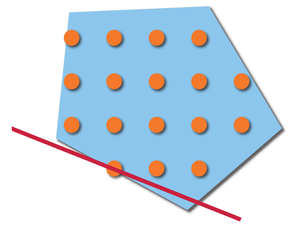
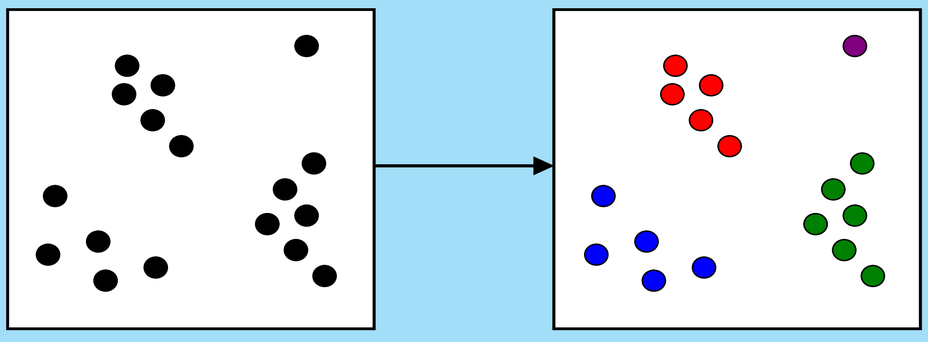
Clusteranalyse



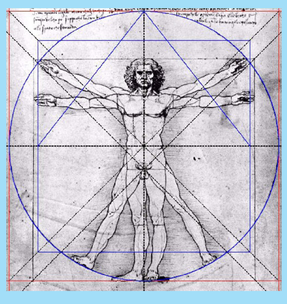
Was ist Clusteranalyse?

Clusteranalyse teilt eine Menge von Datenpunkten in Gruppen (Cluster) ein. Die erhaltenen Gruppen sollen die zugrunde liegende natürliche Struktur der Daten wiedergeben.



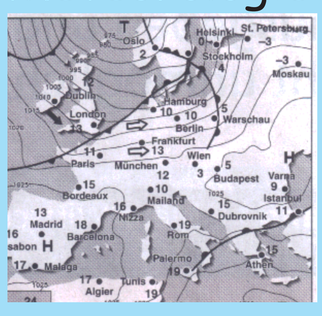
Wo wird Clusteranalyse angewandt?

Medizin



Internet

Meteorologie



Abstands- und Ähnlichkeitsmaße

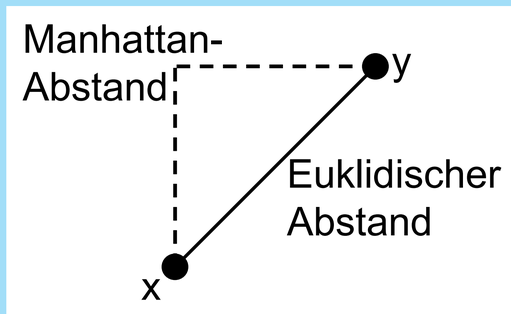
Um Objekte zu clustern, werden sie über Abstands- bzw. Ähnlichkeitsmaße miteinander verglichen.

1. Euklidischer Abstand:

$$d_{euklidisch}(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

2. Manhattan-Abstand:

$$d_{Manhattan}(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$



3. Jaccard-Metrik:

$$d_{Jaccard}(A,B)=1-\frac{|A\cap B|}{|A\cup B|}$$

Beispiel: „Lotto", „Otto", „Eisdiele"

$$d_{Jaccard}(Otto,Lotto)=1-\frac{|\{o,t\}\cap\{l,o,t\}|}{|\{o,t\}\cup\{l,o,t\}|}=1-\frac{|\{o,t\}|}{|\{l,o,t\}|}=1-\frac{2}{3}=\frac{1}{3}$$

$$d_{Jaccard}(Otto,Eisdiele)=1-\frac{|\{\ \}|}{|\{o,t,e,i,s,d,l\}|}=1$$

$$d_{Jaccard}(Lotto,Eisdiele)=1-\frac{|\{l\}|}{|\{l,o,t,e,i,s,d\}|}=\frac{6}{7}$$

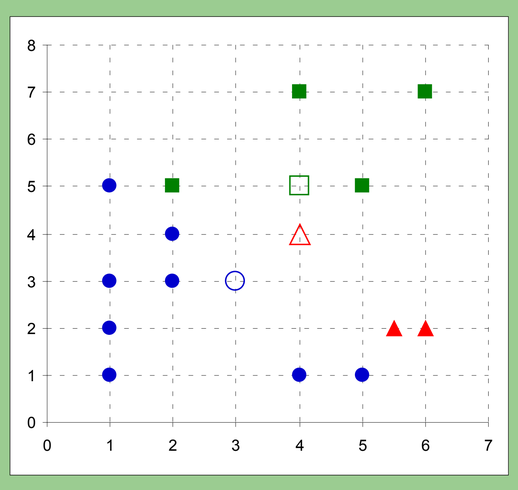
K-means-Algorithmus

Ziel: Einteilung der Daten in K Cluster

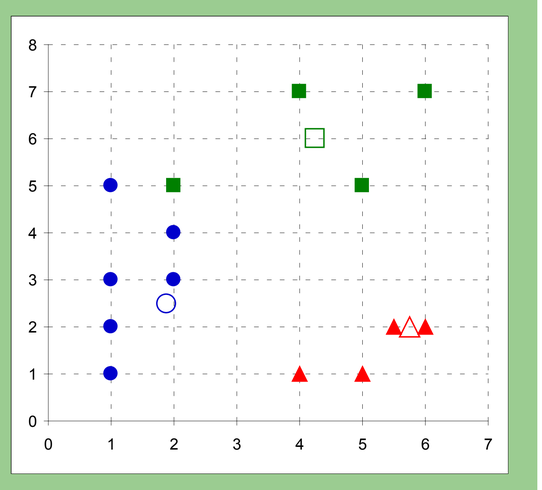
Vorgehen:
1. Wähle K Punkte als Anfangszentren
   wiederhole
2. Ordne jeden Punkt dem Cluster zu, zu dessen Zentrum er am nähesten ist
3. Berechne alle Clusterzentren neu
   bis sich die Zentren nicht mehr ändern

Beispiel:

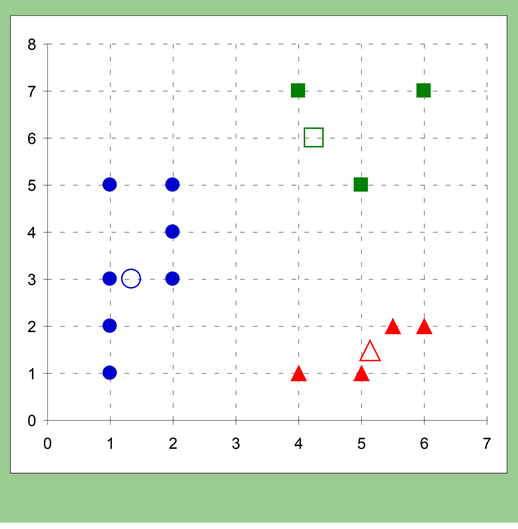
(a)



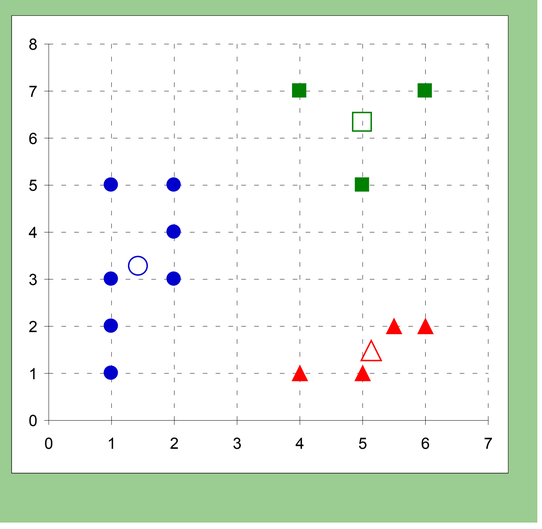
(b)



(c)



(d)

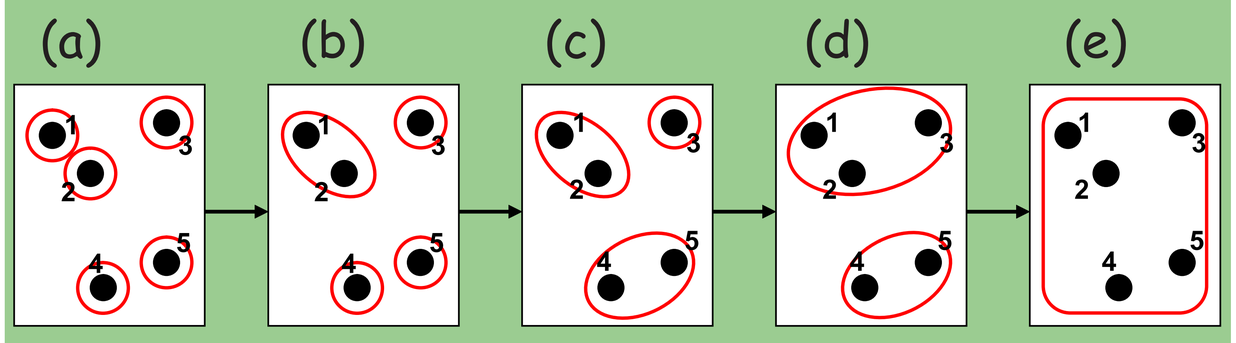


Agglomerierendes Hierarchisches Clustern

Ziel: Bildung einer hierarchischen Folge von Clustern

Vorgehen:
1. Weise jedem Punkt ein Cluster zu
   wiederhole
2. Berechne die Abstände zwischen je zwei Clustern
3. Vereinige die beiden Cluster mit dem geringsten Abstand
   bis alle Daten in einem Cluster sind

Beispiel:



Die entstehende Baumstruktur kann man in einem **Dendrogramm** visualisieren.

Damit das Endergebnis nicht ein Cluster mit allen Daten ist, würde man im obigen Beispiele den Algorithmus im Schritt (d) abbrechen.