Föreläsning 7 - Klusteranalys

Josef Wilzen

2021-09-20

Outline

Intro

- Oövervakad inlärning
  - **Klusteranalys**
  - **Associationsanalys och sekventiella mönster**
  - Dimensionality reduction techniques
  - PCA, Faktormodeller

Intro

- Målet med klusteranalys är att dela upp datamaterialet i grupper (kluster) som är intressanta och/eller användbara
- Vi vet inte i förväg vilka grupper som kommer att blidas
- Ingen responsvariabel

Klusteranalys

1. Ge exempel på områden där klusteranalys kan vara användbart
2. Hur många kluster finns i bilden nedan?

Definition av kluster

Begreppet ”kluster” är inte entydigt definierat.

- Tillämpningsområden
  - Biologi (taxonomi/gener)
  - Informationssökning (sökmotorer)
  - Psykologi och medicin
  - Kunddata
  - Sociala medier/nätverk

Klassificering och klustering

- Klassificeringsmetoder som beskrevs tidigare är exempel på övervakad klassificering – markerar nya objekt, utgår från originaldata (data, etiketter)
- Klusteranalys är ett exempel på **oövervakad** klassificering – härleder en etikett för objekt, utgår endast från data

Klustringstyper

- **Partitionell**: data är indelad i ett antal oöverlappande kluster,
  eller
- **Hierarkisk**: delkluster är tillåtna, kluster är representerade som ett träd

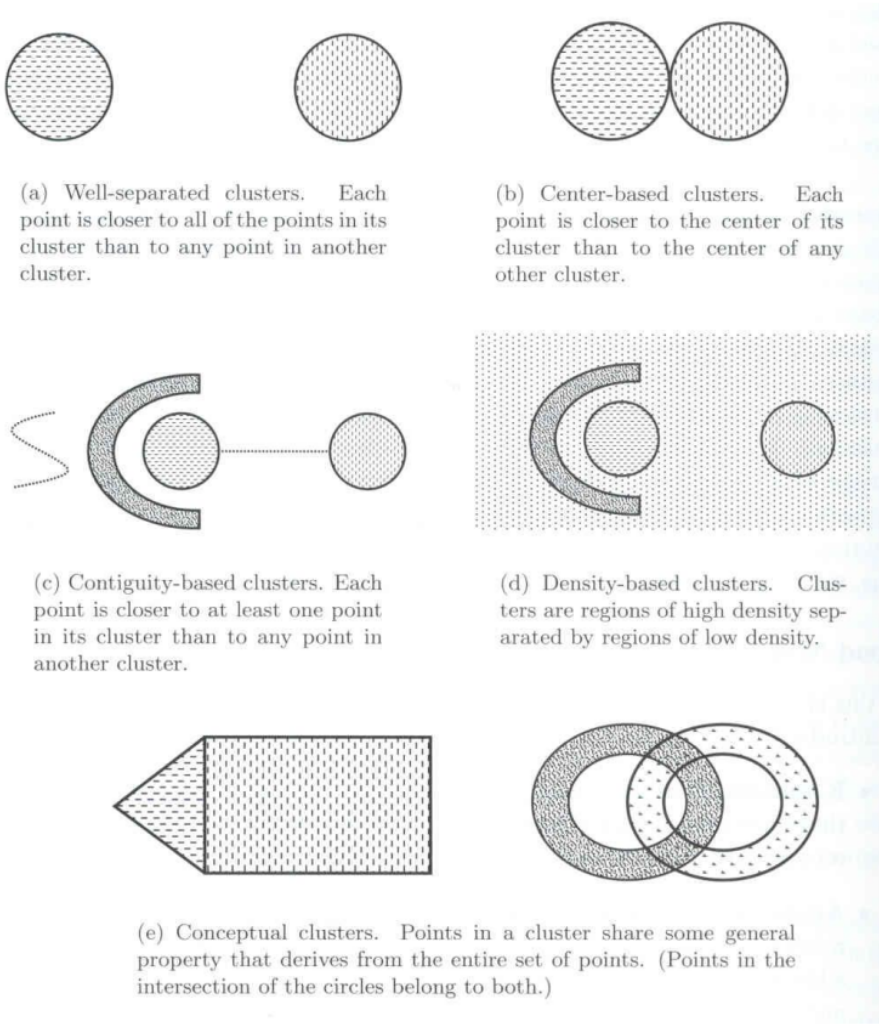
- **Uteslutande**: ett objekt tillhör ett kluster,
  eller
- **Överlappande**: ett objekt hör till några kluster,
  eller
- **Fuzzy**: Ett objekt hör till olika kluster med en specifik sannolikhet

Klustringstyper

- **Fullständig**: varje objekt tillskrivet ett kluster,
  eller
- **Ofullständig**: somliga objekt är inte tillskrivna något kluster

Klustertyper

- Separerade
- Angränsande/intilliggande
- Centroid- eller prototypbaserade
- Densitet- eller täthetsbaserade
- Konceptuella



(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.

(d) Density-based clusters. Clusters are regions of high density separated by regions of low density.

(e) Conceptual clusters. Points in a cluster share some general property that derives from the entire set of points. (Points in the intersection of the circles belong to both.)

K-means klustring

K-means klustring

- Centroid-baserad och partitionell klustringsmetod
  - Centroid = en punkt som ska representera/sammanfatta alla obs i ett kluster
- Enkel och ofta effektiv metod
- K: hyperparameter, antalet kluster

Algorithm 8.1 Basic K-means algorithm.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

K-means klustring

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

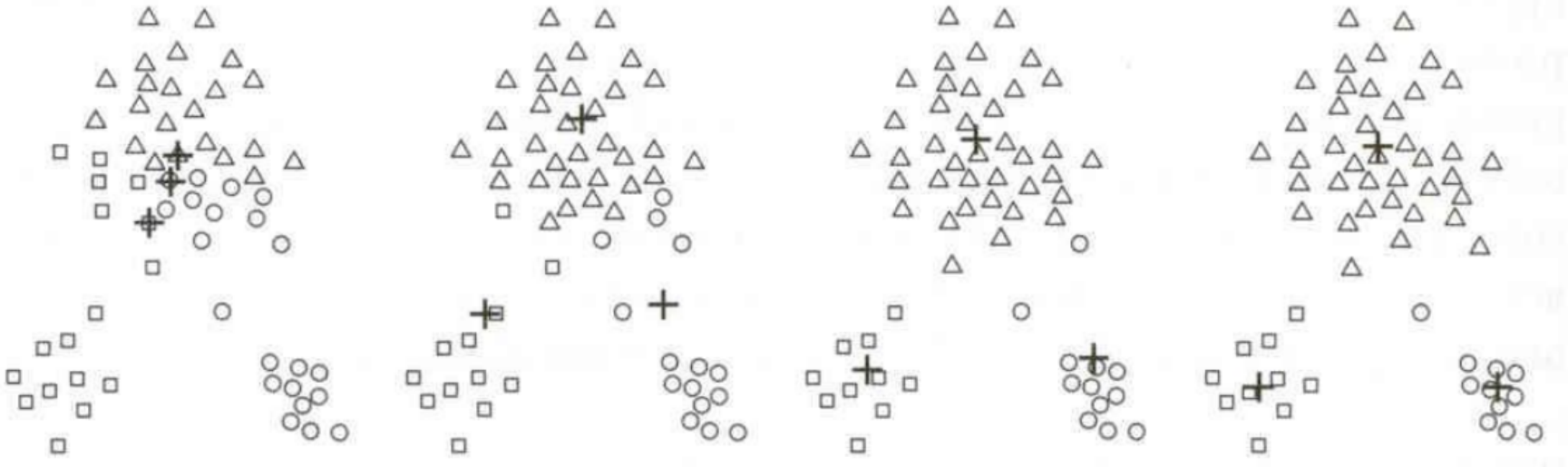
2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Exempel



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

K-means

- Låt c_i vara centroider för kluster i, låt C_i vara en mängd med alla obs i kluster i
- Vi behöver definera ett avståndsmått
  - Används för att mäta avstånd mellan c_i och övriga obs
  - Vanligast är euklidiskt avstånd: låt p och q vara två vektorer

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

  - Notera: $d(p,p) = \sqrt{\sum_{i=1}^{n} p_i^2} = \sqrt{p^T p} = ||p||$ är den euklidiska normen. Ridge använder $||p||^2 = \sum_{i=1}^{n} p_i^2$

K-means

- K-means minimerar *SSE*
- SSE i ett kluster:

$$E_{c_i} = \sum_{x \in C_i} d(x, c_i)^2$$

- Totala SSE för alla kluster

$$SSE = \sum_{i=1}^{K} E_{c_i} = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, c_i)^2$$

- I det euklidiska rummet beräknas centroider som

$$c_i = \frac{1}{n_i} \sum_{x \in C_i} x$$

- K-means algoritmen hittar en ett **lokalt** minima

Exempel

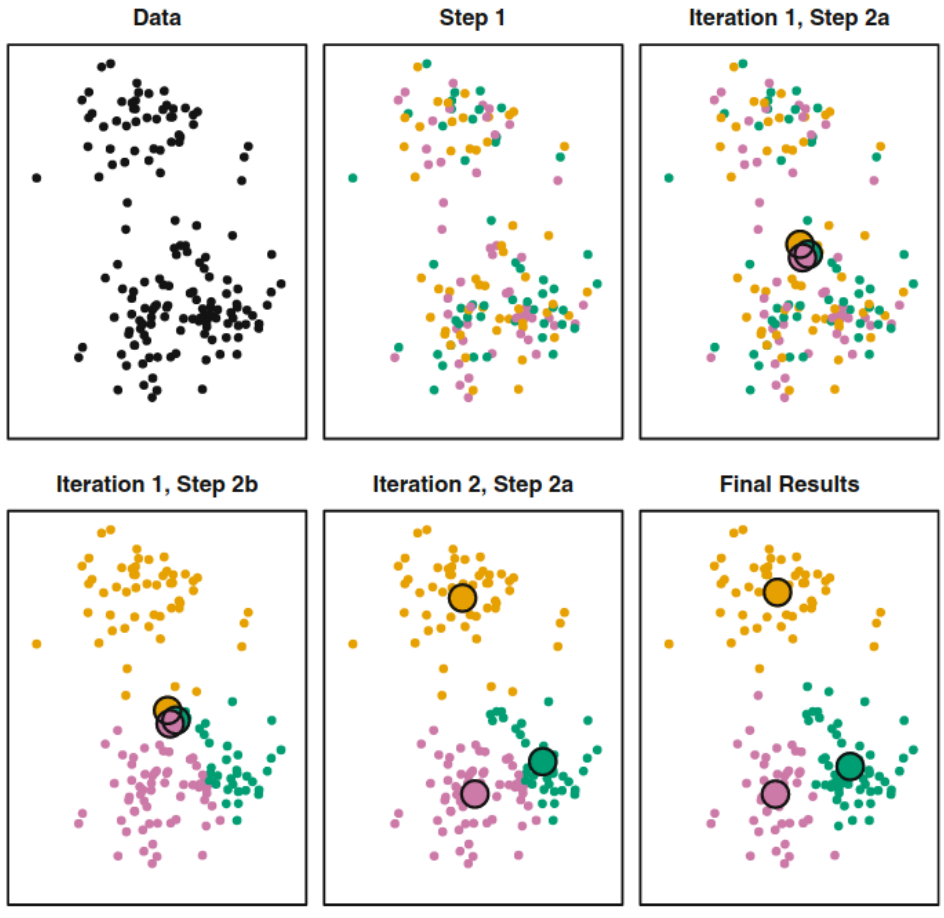
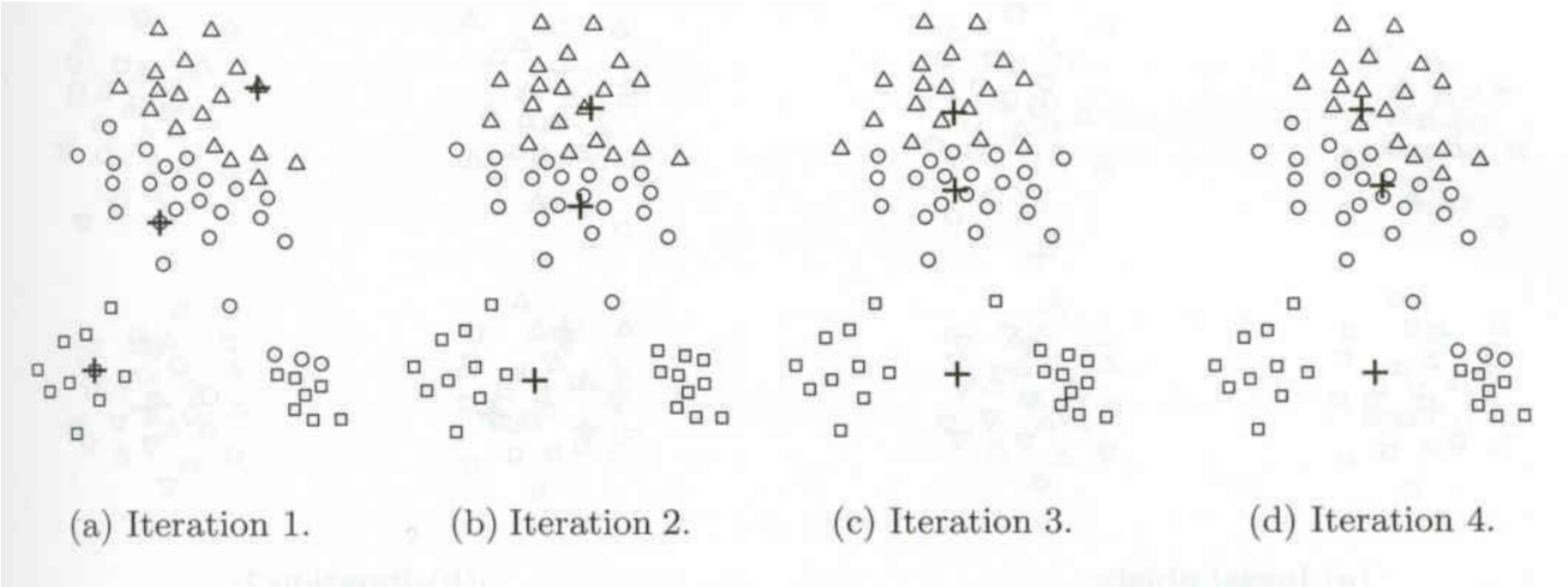


FIGURE 10.6. *The progress of the K-means algorithm on the example of Figure 10.5 with K=3.* Top left: *the observations are shown.* Top center: *in Step 1 of the algorithm, each observation is randomly assigned to a cluster.* Top right: *in Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.* Bottom left: *in Step 2(b), each observation is assigned to the nearest centroid.* Bottom center: *Step 2(a) is once again performed, leading to new cluster centroids.* Bottom right: *the results obtained after ten iterations.*

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Startvärden

- Vi måste välja startvärden för centroiderna, valet påverkar starkt utgången av algoritmen
- Exempelvis, om man väljer dåliga startpunkter:

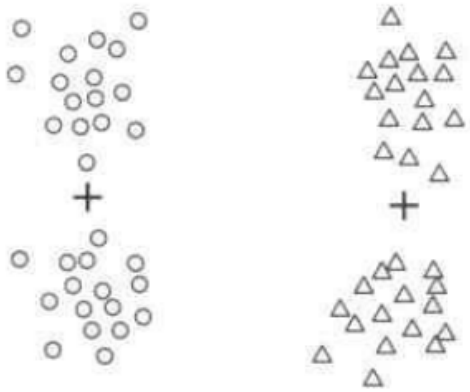


- Vanlig metod är att köra algoritmen många gånger med olika slumpmässiga startvärden.

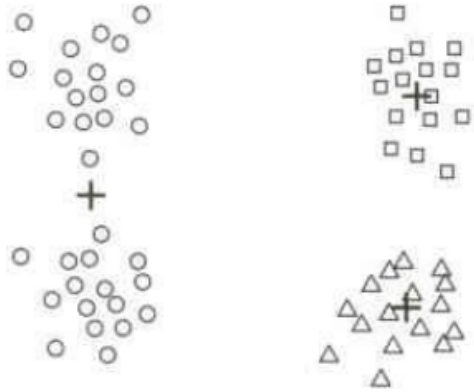
Halverande K-means

- Algoritm som motverkar problemet med val av start-centroider
- Dela upp datamängden i två kluster, välj ett och dela upp i två, välj ett (av de nuvarande 3) och dela upp osv...
  - Valet av kluster kan göras med avseende på flest observationer, störst SSE eller annat kriterie
- Uppdelningen kan liknas vid ett binärt träd

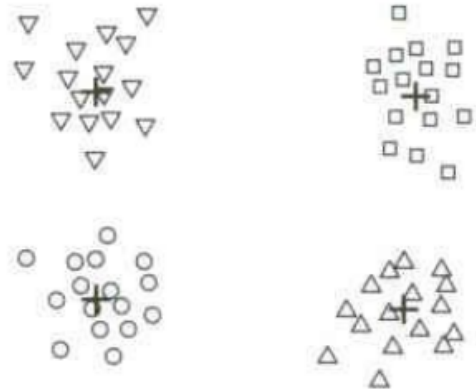
Halverande K-means



(a) Iteration 1.



(b) Iteration 2.



(c) Iteration 3.

Halverande K-means

Algorithm 8.2 Bisecting K-means algorithm.

1: Initialize the list of clusters to contain the cluster consisting of all points.
2: **repeat**
3: Remove a cluster from the list of clusters.
4: {Perform several "trial" bisections of the chosen cluster.}
5: **for** $i = 1$ to *number of trials* **do**
6: Bisect the selected cluster using basic K-means.
7: **end for**
8: Select the two clusters from the bisection with the lowest total SSE.
9: Add these two clusters to the list of clusters.
10: **until** Until the list of clusters contains K clusters.

K-means++

- Algoritm som motverkar problemet med val av start-centroider
- Kostnad för vanlig k-means: $O(n \cdot K \cdot d \cdot i)$, för många dataset $\approx O(n)$
  - n: antal obs, K: antal kluster, d: antalet variabler, i: antalet iterationer till konvergens
  - Ordo eller Big O notation, se här.
- SSE kan bli godtyckligt dåligt med k-means

K-means++

1. Välj en centroid uniformt slumpmässsigt från observationerna
2. För varje datapunkt x, beräkna avståndet $d(x, c_i)$ mellan x och den närmaste centroiden c_i som redan har valts.
3. Välj en datapunkt som centroid genom att:
    1. slumpa en punkt (som inte redan är en centroid), med hjälp av viktade sannolikheter, där vikterna är proportionella mot $d(x, c_i)^2$
4. Upprepa steg 2) och 3) tills K centroider har valts
5. Givet de valda centroiderna: kör vanlig k-means klustering

K-means++

- Generellt: k-means++ förbättrar slutgiltiga SSE mycket
- Steg 1)-4) tar extra tid att beräkna, men sen krävs det ofta mycket mindre iterationer innan vanliga k-means konvergerar i 5)
  - Vanligt att k-means++ är dubbelt så snabb som k-means med avseende på total beräkningstid

K-means: kommentarer

- Enkel och ganska effektiv
- Känslig mot initialiseringsproblem
  - Halverande k-means, k-means++
- Skapar kluster som är klotformade i $\mathbb{R}^d$ och är är linjärt spearerade
  - Andra former fungerar sämre
- Ger en centroid/prototyp för varje kluster: kan användas för att beskriva klustren
- Har svårt att identifiera kluster av olika storlekar eller med olika tätheter
- Känslig mot extremvärden

K-means: utökningar

- Kernel k-means: kan forma kluster av olika former, med icke-linjära speareringsgränser
- Gaussian mixture models/clustering:
  - Varje kluster beskrivs med en multivariat normalfördelning
  - Skattas med expectation–maximization (EM) algorithm
- k-medoids/Partitioning Around Medoids (PAM): använder medioder som center
- k-medians clustering: använder medianer istället

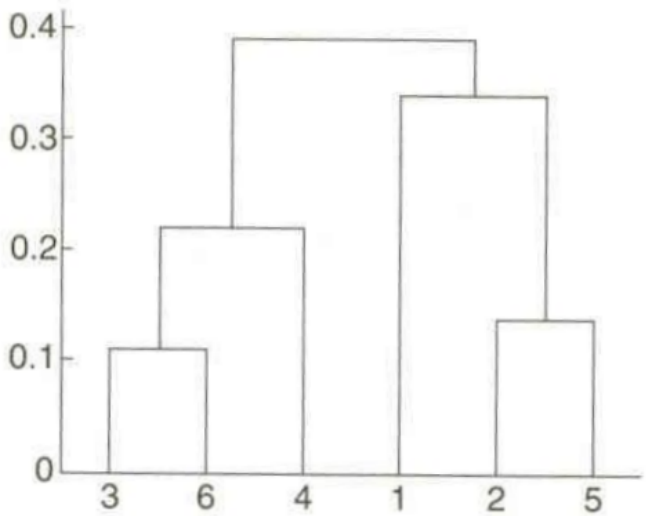
Hierarkisk klustring

Hierarkisk klustring

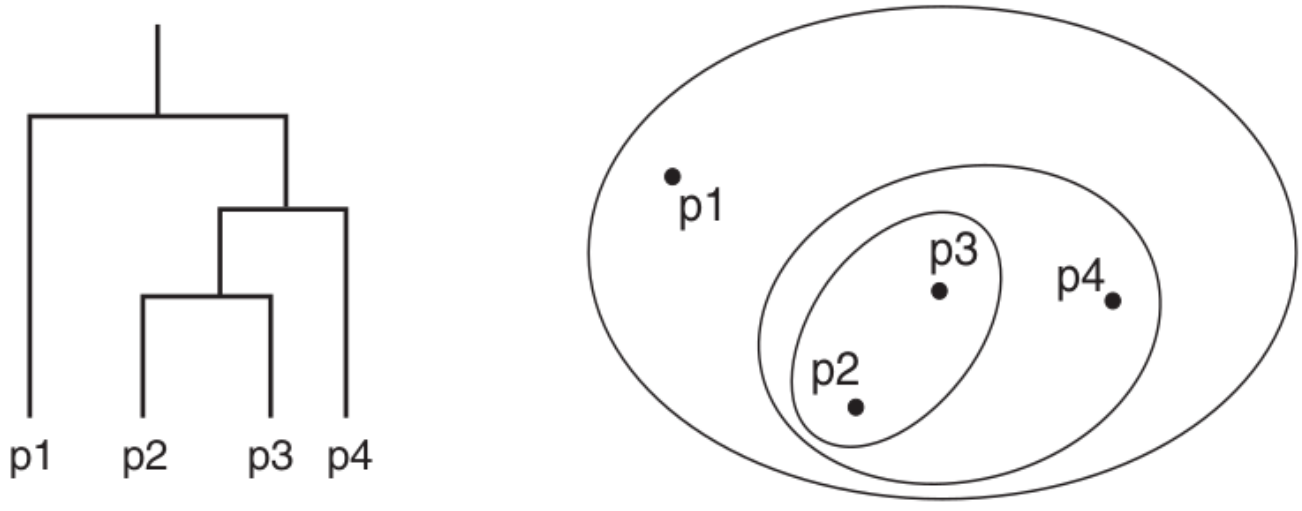
- Två typer:
  - **Agglomerativ hierarkisk klustring**
  - Diversiv hierarkisk klustring
- Skapar en hierarki med kluster
  - Subkluster som har subkluster, som har subkluster...

Agglomerativ hierarkisk klustring

- Börja med enstaka observationer som individuella kluster och slå ihop närmaste par av kluster steg för steg. Detta upprepas tills alla observationer är i ett kluster.
- Processen visualiseras i ett s.k. dendrogram
  - Vågrät axel innehåller observationsnummer (notera att ordningen här är godtycklig)
  - Lodrät axel mäter avstånd mellan kluster
  - Förgreningen mäter vilka kluster och vid vilket avstånd dessa slås ihop



Agglomerativ hierarkisk klustring



(a) Dendrogram. (b) Nested cluster diagram.

Figure 7.13. A hierarchical clustering of four points shown as a dendrogram and as nested clusters.

Källa: Introduction to Data Mining

Dendrogram

- Dendrogrammet visar *alla* ihopslagningar
- Vi måste manuellt ange när vi anser ihopslagningarna ska sluta: Hur många kluster?
  - Subjektivt
  - När avstånden mellan ihopslagningar (lodräta linjer) är ”nog stort”

Dendrogram

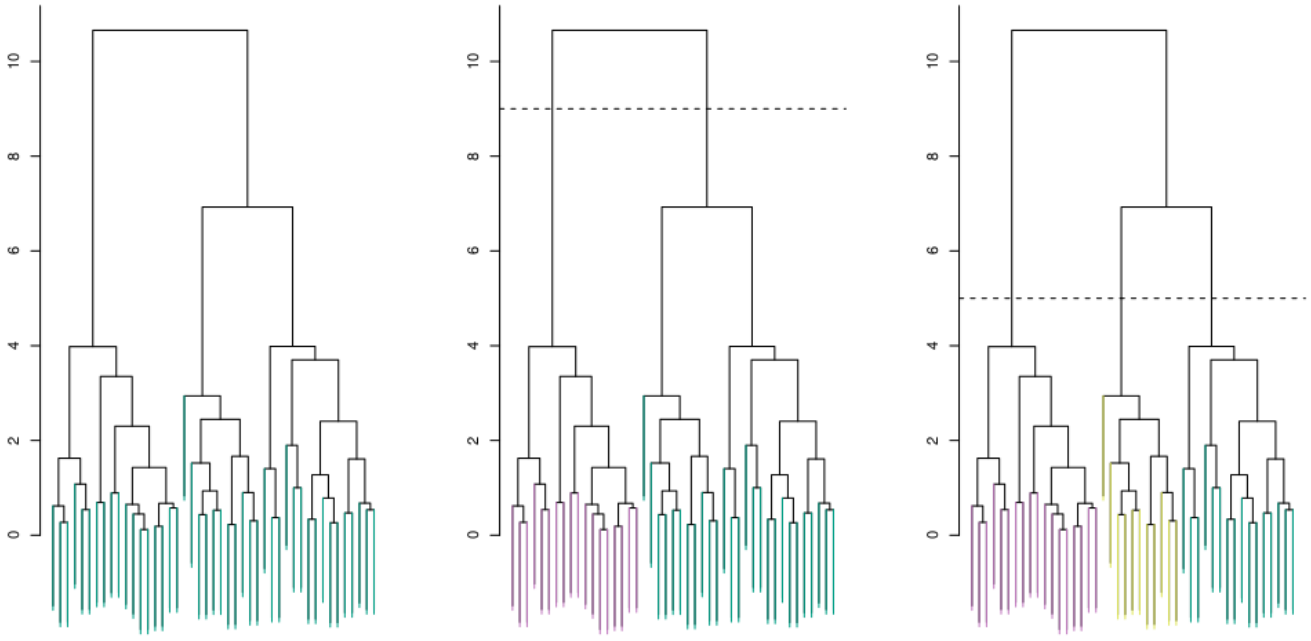


FIGURE 12.11. Left: *dendrogram obtained from hierarchically clustering the data from Figure 12.10 with complete linkage and Euclidean distance.* Center: *the dendrogram from the left-hand panel, cut at a height of nine (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.* Right: *the dendrogram from the left-hand panel, now cut at a height of five. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure.*

Källa: An Introduction to Statistical Learning

Algoritm

Algorithm 7.4 Basic agglomerative hierarchical clustering algorithm.

1: Compute the proximity matrix, if necessary.
2: **repeat**
3: Merge the closest two clusters.
4: Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
5: **until** Only one cluster remains.

Källa: Introduction to Data Mining

Proximity matrix är en matris innehållande närheten mellan kluster, även distansmatriser kan användas

Beräkning av avstånd mellan två kluster

Då kluster ofta innehåller flera observationer behövs en metod för att definiera hur avstånd beräknas, även kallad *länkningsmetod*.
Låt C_i och C_j vara två kluster.

- MIN eller Single (enkel länkning):

$$prox(C_i, C_j) = \min_{x \in C_i, y \in C_j} dist(x, y)$$

- MAX eller Complete (fullständig länkning):

$$prox(C_i, C_j) = \max_{x \in C_i, y \in C_j} dist(x, y)$$

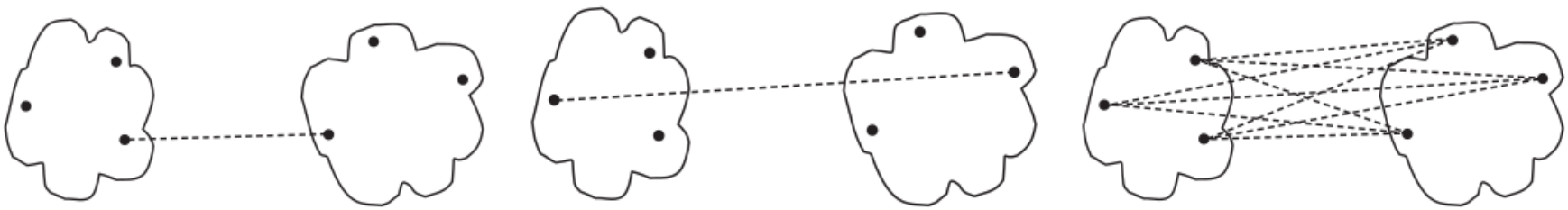
Beräkning av avstånd mellan två kluster

- Group average (genomsnitts länkning):

$$prox\left(C_i, C_j\right) = \frac{1}{\left(n_i \cdot n_j\right)} \sum_{x \in C_i, y \in C_j} dist\left(x, y\right)$$

  där n_i och n_j är antalet obs i kluster i och j respektive.
- Wards/Centroid metod: närhet defineras som hur mycket kvadrerade fel ökar när två kluster slås ihop. Samma kostandsfunktion som k-means.

Beräkning av avstånd mellan två kluster



(a) MIN (single link). (b) MAX (complete link). (c) Group average.

Källa: Introduction to Data Mining

Egenskaper

- Ingen global funktion att optimera (jämf. K-means)
- Group average- och olika centroid metoder kan ta hänsyn till olika klusterstorlekar när ett par kluster förenas
- Ihopslagningar är slutgiltiga och går inte att ta isär
- Närhetsmåttet kan påverka resultatet
  - extremvärden eller brus
- Passar bra för data som har en hierarkisk struktur

Egenskaper

- Minneskomplexitet: $O\left(n^2\right)$
- Tidskomplexitet: $O\left(n^3\right)$, med smarta datastrukturer $O\left(n^2 log\left(n\right)\right)$

Exempel

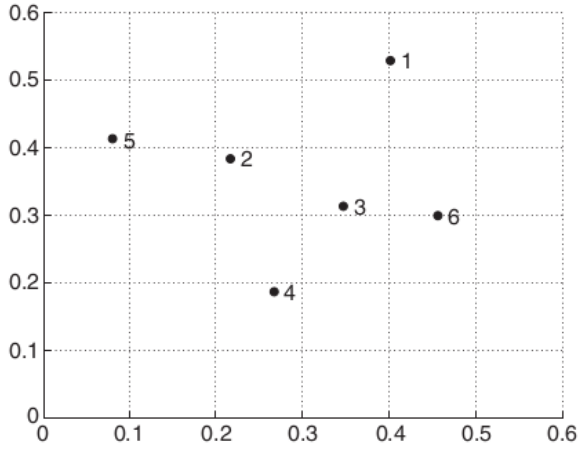


Figure 7.15. Set of six two-dimensional points.

Point	x Coordinate	y Coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

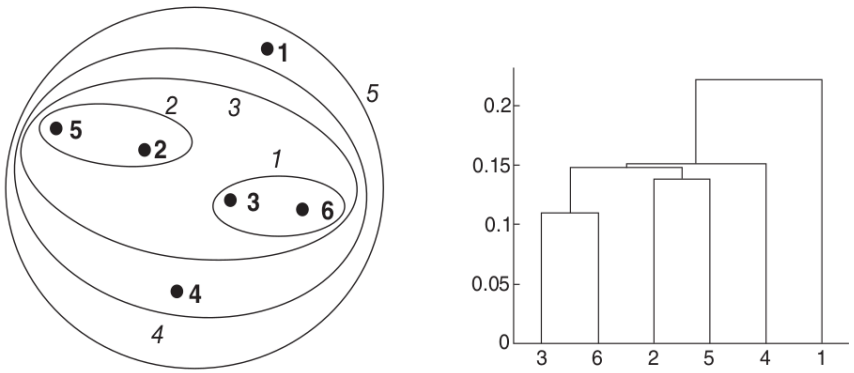
Table 7.3. xy-coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Table 7.4. Euclidean distance matrix for six points.

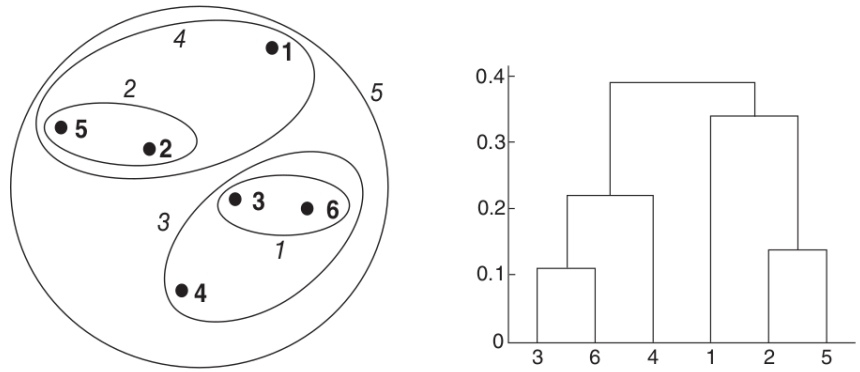
Källa: Introduction to Data Mining

Exempel



(a) Single link clustering. (b) Single link dendrogram.

Figure 7.16. Single link clustering of the six points shown in Figure 7.15.

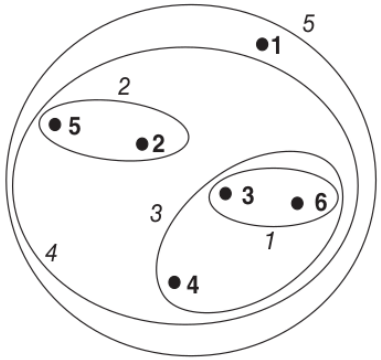


(a) Complete link clustering. (b) Complete link dendrogram.

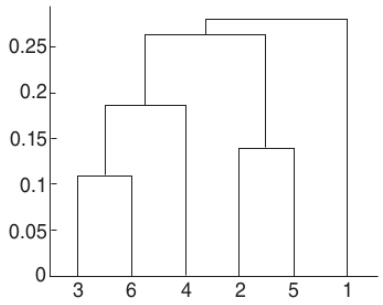
Figure 7.17. Complete link clustering of the six points shown in Figure 7.15.

Källa: Introduction to Data Mining

Exempel

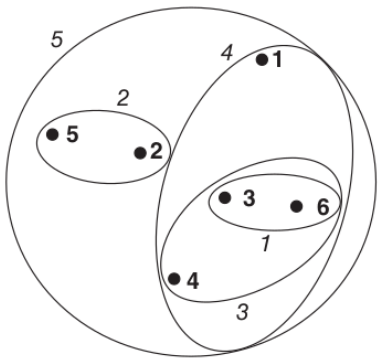


(a) Group average clustering.

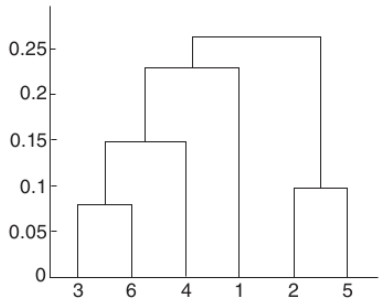


(b) Group average dendrogram.

Figure 7.18. Group average clustering of the six points shown in Figure 7.15.



(a) Ward's clustering.



(b) Ward's dendrogram.

Figure 7.19. Ward's clustering of the six points shown in Figure 7.15.

Källa: Introduction to Data Mining

Avslut

- Kurshemsidan
- Labben