LINKÖPING UNIVERSITY

732G12 Data Mining

Föreläsning 9

Josef Wilzén
IDA, Linköping University, Sweden

Dagens föreläsning

- Introduktion
- K-means klustring
- Hierarkisk klustring

Projektet

- Info finns på kurshemsidan
- Börja kolla på uppgiften och hitta data!

Introduktion

- Oövervakad inlärning: Lära sig data **utan** responsvariabel!
- Flera olika algoritmer:
  - **Klusteranalys**
  - Associationsanalys
  - Sekventiella mönster
  - Dimensionalty reduction techniques
  - PCA, Faktormodeller
  - Representation Learning

Introduktion

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 bildas.

Ingen responsvariabel.

Om erhållna kluster är intressanta och/eller användbara beror på kontexten/problemet/området.

Klusteranalys

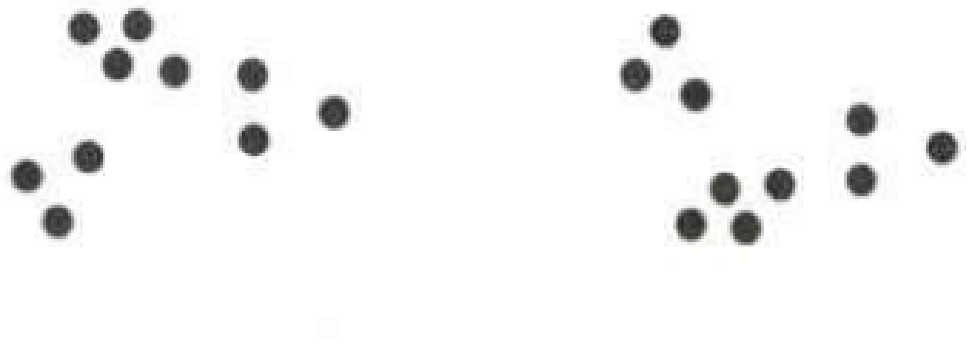
Ta 5 minuter att fundera på följande frågor:

1. Hur många kluster finns det i bilden?

2. Kom på något område där klusteranalys kan vara användbart.

Klusteranalys

Ett ”kluster” är inte entydigt definerat.



Tillämpningsområden:

- Biologi (toxonomi/gener)
- Informationssökning (Sökmotorer)
- Psykologi och medicin
- Kunddata
- Sociala medier/nätverk

Klassificering och klustering

- Klassificeringsmetoder som vi jobbat med tidigare är exempel på övervakad inlärning. Ger etiketter till nya objekt, utgår från orginaldata som har etiketter.
- Klusteranalys är ett exempel på oövervakad inlärning - vi härleder en etikett för objekt, utgår endast från data.

Klustringstyper

Partionell: Data är indelad i ett antal oöverlappande kluster.

Hierarkisk: Delkluster är tillåtna, kluster är representerade som ett träd.

Uteslutande: Ett objekt tillhör ett kluster.

Överlappande: Ett objekt hör till några kluster.

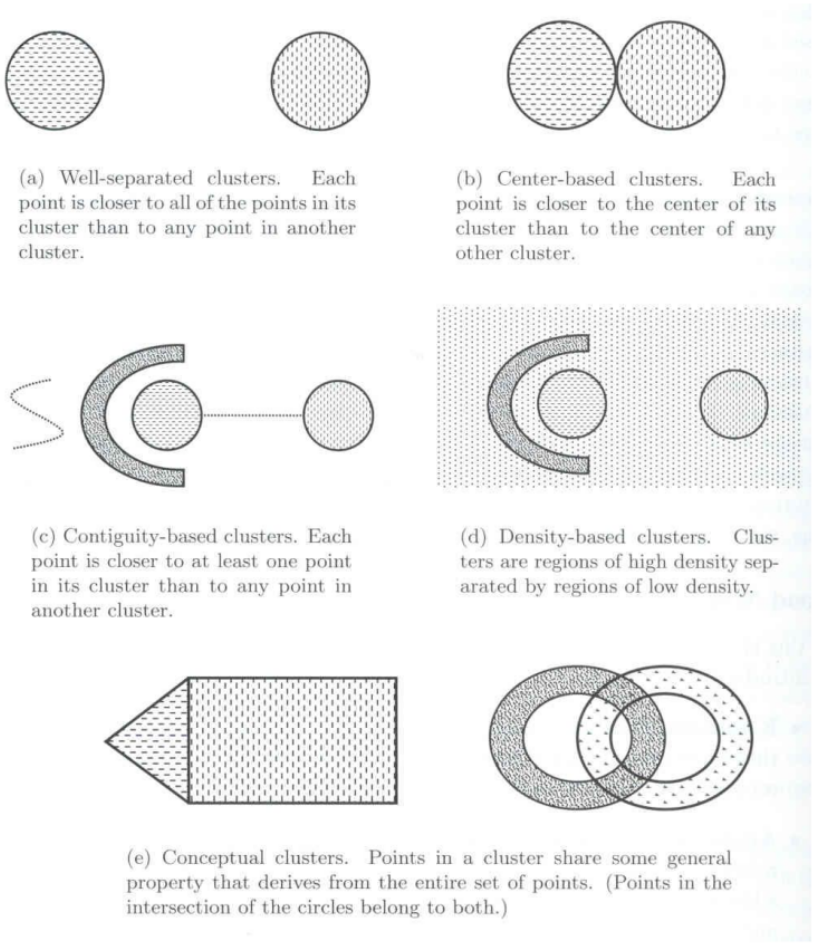
Fuzzy: Ett objekt hör till olika kluster med en specifik sannolikhet.

Fullständig: Varje objekt är tillskrivet (minst) ett kluster.

Ofullständigt: Vissa objekt är inte tillskrivna något kluster.

Klustertyper

- Separerade
- Angränsande/intilliggande
- Centroid- eler 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

- Centroid-baserad och partionell klustringsmetod.
  - Centroid är en punkt som ska representera/sammanfatta alla observationer i ett kluster.
- Enkel och ofta effektiv metod.
- K: hyperparameter, antalet klasser.

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.

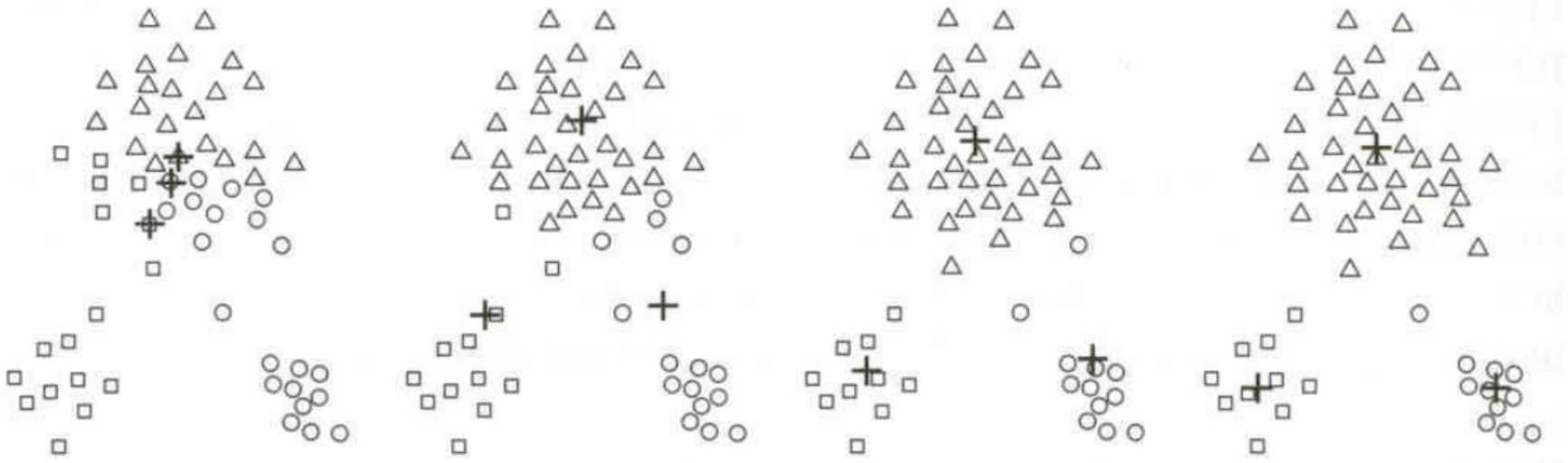
K-means klustring

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

K-means klustring: Exempel



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

K-means klustring: Detaljer

- Låt c_i vara centroid för kluster i. Låt C_i vara en mängd med alla observationer i kluster i.
- Vi behöver ett avståndsmått
  - Används för att mäta avståndet mellan c_i och övriga observationer.
  - Vanligast är euklidiskt avstånd,

$$d(x, x') = \sqrt{\sum_{i=1}^{p} (x_i - x'_i)^2}$$

  - Finns såklart många andra val som kan göras.

K-means klustring: Detaljer

- K-means minimerar SSE
- SSE i ett kluster ges av

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

- Totala SSE för alla kluster

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

- I det euklidiska rummet beräknas centroider som

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

- K-means algoritmen hittar ett lokalt minima.

K-means klustring: Exempel

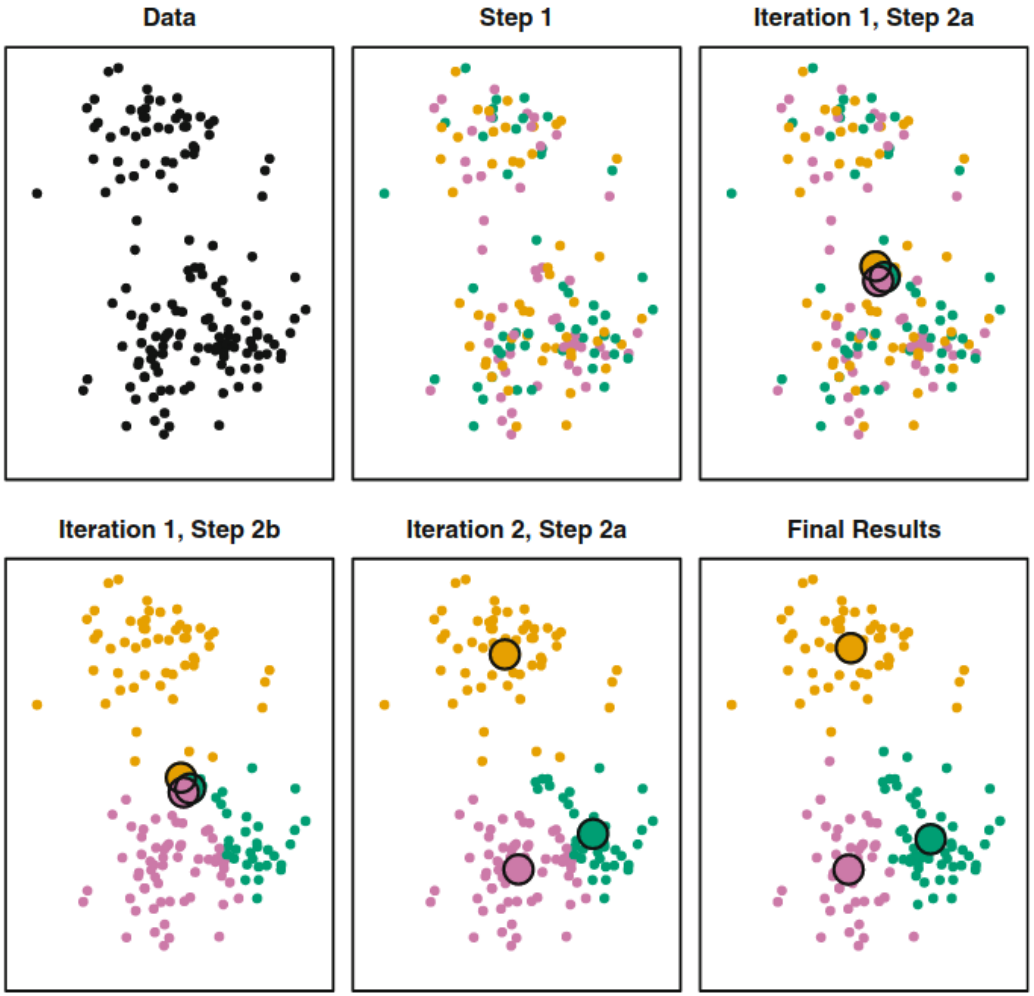
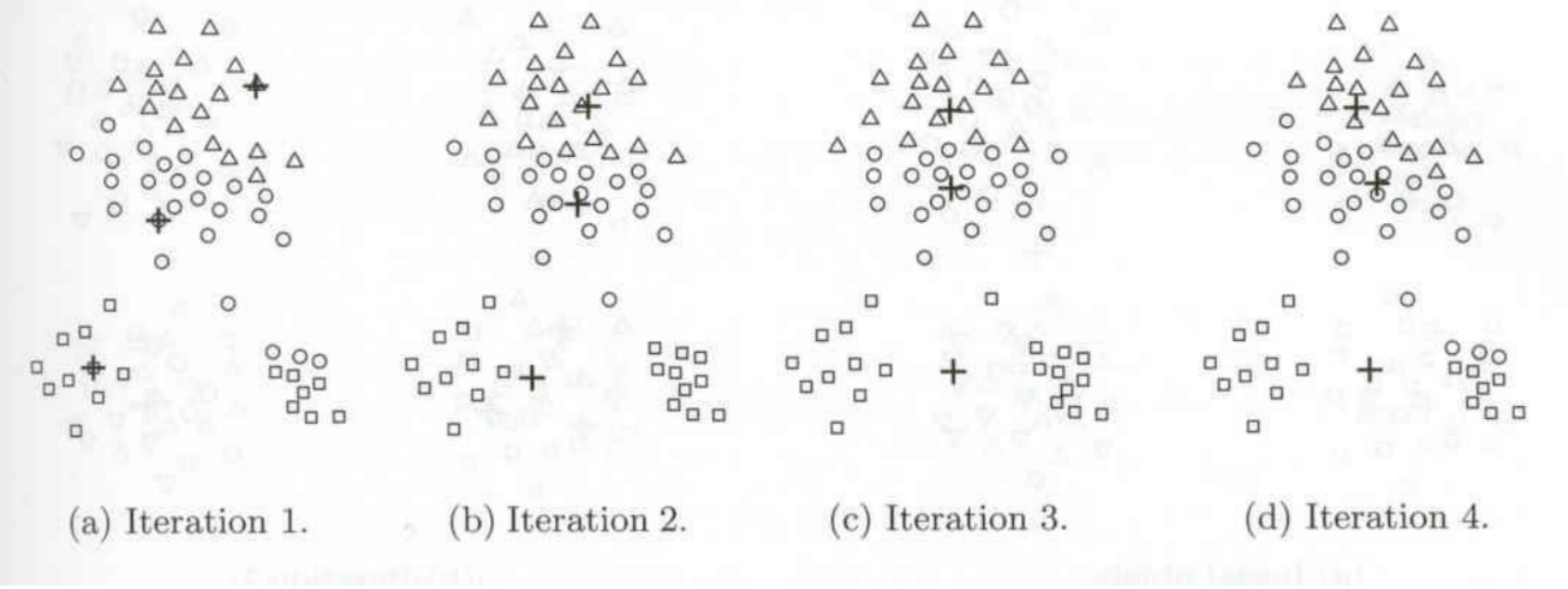


FIGURE 10.6. *The progress of the K-means algorithm on the example of Figure 10.5 with K=3.* Top left: *the observations are shown.* Top center: *in Step 1 of the algorithm, each observation is randomly assigned to a cluster.* Top right: *in Step 2(a), the cluster centroids are computed. These are shown 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*

K-means klustring: Startvärden

Vi måste välja våra initiala gissningar för centroider. Detta val påverkar starkt utgången av algoritmen.

Dåliga startvärden kan ge dåliga resultat.

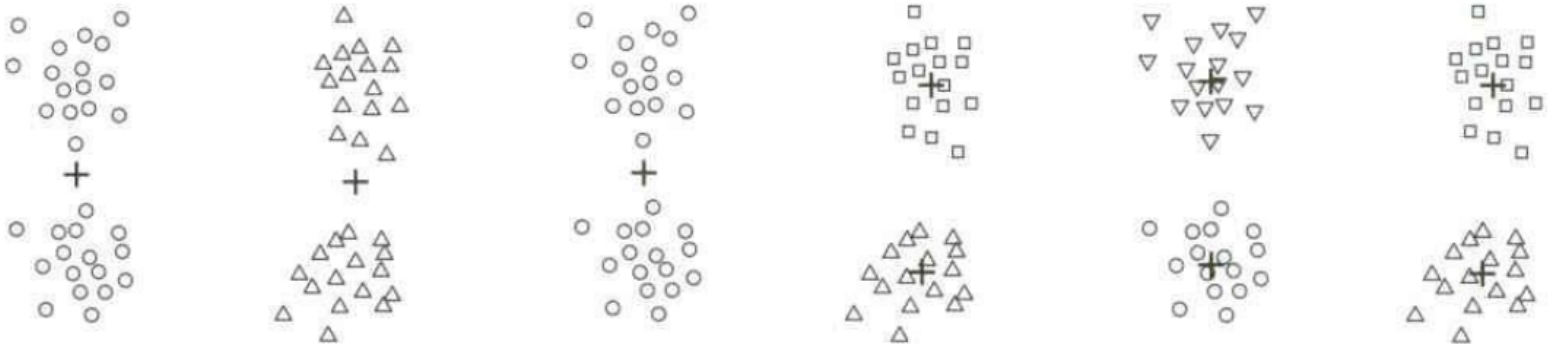


Vanlig metod är att köra algoritmen många gångenr med olika (slumpade) startvärden.

Halverande K-means

- Algoritm som motverkar problemet med startvärden.
- Dela upp datamängden i två kluster, välj ett och dela upp det i två osv.
  - Valet av kluster kan göras med avseende på flest observation, 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 för att hitta startvärden till K-means.

1. Välj en centroid uniformt slumpmässigt från observationerna.
2. För varje datapunkt x beräkna avstånder $d(x, c_i)$ mellan x och den närmaste centroiden (som redan valts).
3. Välj en datapunkt som centroid genom att:
   - Slumpa en punkt med hjälp av viktade sannolikheter, där vikterna är proportionella mot $d(x, c_i)^2$.
4. Upprepa 2 och 3 tills K centroider valts ut.
5. Kör vanlig K-means med dessa startcentroider.

K-means++

- Generellt: K-means++ förbättrar SSE mycket över slumpade startvärden.
- Tar extra tid att bestämma startvärden, men K-means konvergerar mycket snabbare än med slumpade startvärden.
- Vanligt att K-means++ är dubbelt så snabb som K-means med slumpade startvärden.

K-means klustring: Kommentarer

- Enkel och ganska effektiv.
- Känslig mot startvärden.
  - Kör många gånger med olika startvärden.
  - Halverande K-means.
  - K-means++
- Skapar klotformade kluster och är linjärt separerade.
  - Om datas "naturliga kluster" har andra former fungerar k-means sämre.
- Ger en centroid för varje kluster. Kan användas för att beskriva klustret.
- Har svårt att identifiera kluster av olika storlekar eller med olika tätheter.
- Känslig mot extremvärden.

K-mean klustring: Utökningar

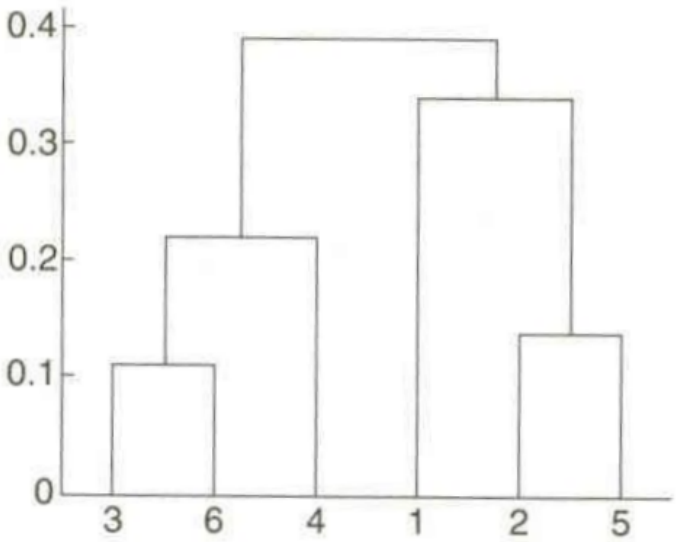
- Kernel K-means: Kan forma kluster av olika former med icke-linjära separationsgränser.
- Gaussian mixture models/klustring:
  - Varje kluster beskrivs med en multivariat normalfördelning.
  - Skattas med expectation-maximization (EM) algoritmen.
- K-medoids/Partitioning Around Medoids. Använder medioder som center (en punkt i datasetet).
- K-medians klustring: använder medianer istället.

Hierarkisk klustring

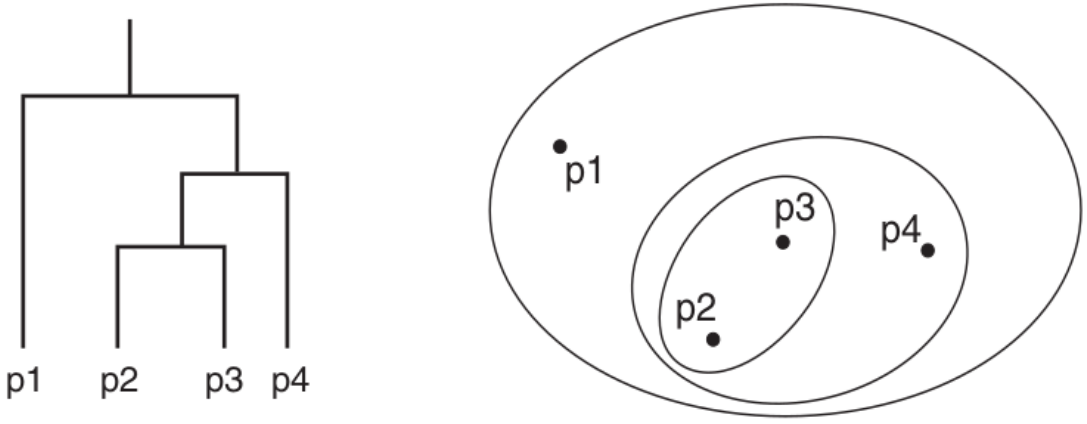
- Två typer:
  - Agglomerativ, bygger underifrån.
  - Diversiv, bygger uppifrån.
- Skapar en hierarki med kluster.
  - Subkluster som har subkluster som har subkluster....

Agglomerativ hierarkisk klustring

- Börja med att ge varje observation sitt egna kluster. Slå ihop närliggande kluster till ett större kluster. Upprepa detta tills alla observationer är i ett kluster.
- Proocessen visualiseras i ett s.k. dendogram.
  - Vågrät axel innehåller observationsnummer (ordningen 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.

Dendogram

- Dendogrammet visar **alla** ihopslagningar.
- Vi måste manuellt ange när vi anser att ihopslagningarna ska sluta: Hur många kluster?
  - Subjektivt
  - När avståndet mellan ihopslagningar är ”stort nog”.

Dendogram

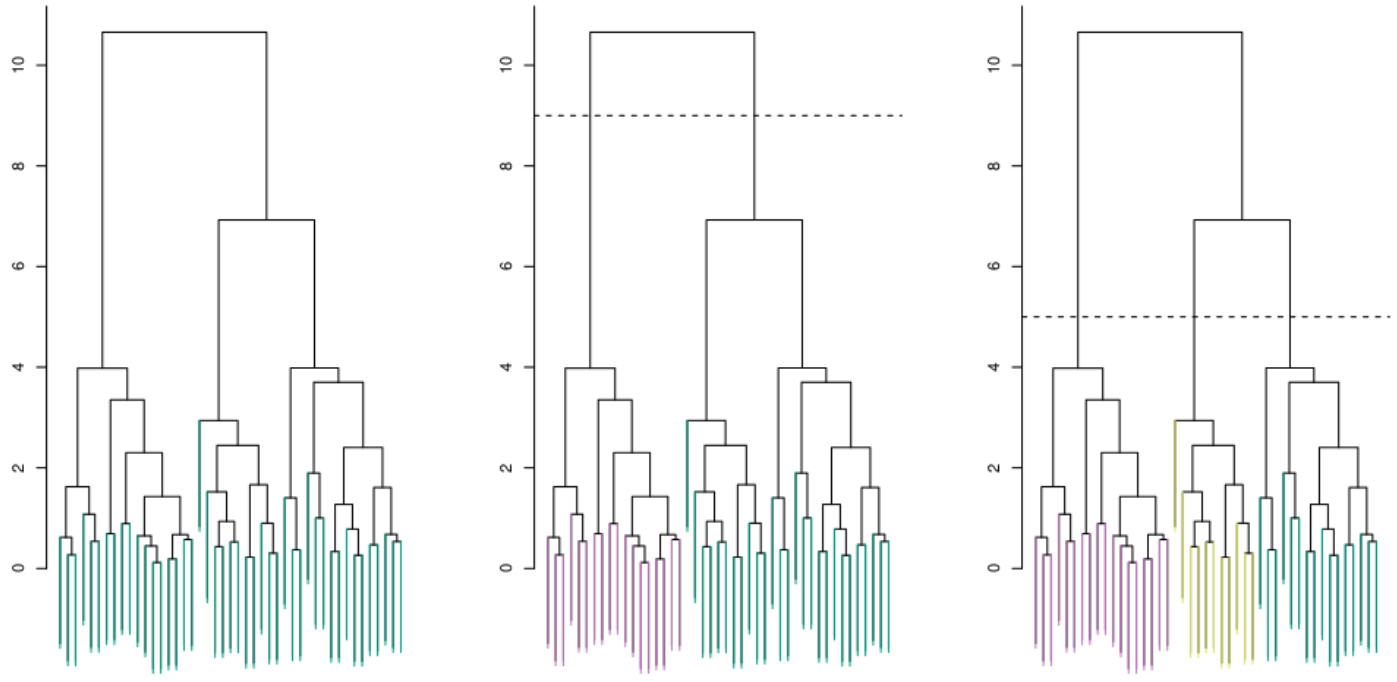


FIGURE 12.11. Left: *dendrogram obtained from hierarchically clustering the data from Figure 12.10 with complete linkage and Euclidean distance.* Center: *the dendrogram from the left-hand panel, cut at a height of nine (indicated by the dashed line). This cut results in two distinct clusters, shown 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.*

Agglomerativ hierarkisk klustring: 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.

Proximity matrix är en matris som innehåller närheten mellan kluster.

Kan använda en distansmatris också.

Beräkning av avstånd mellan två kluster

Då kluster ofta innehåller flera observationer behövs en metod för att definera hur avstånd beräknas, även kallad **länkningsmetod**.

Låt C_i och C_j vara två kluster.

- MIN eller Single linkage (enkel länkning):

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

- MAX eller Complete linkage (fullständig länkning):

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

Beräkning av avstånd mellan två kluster

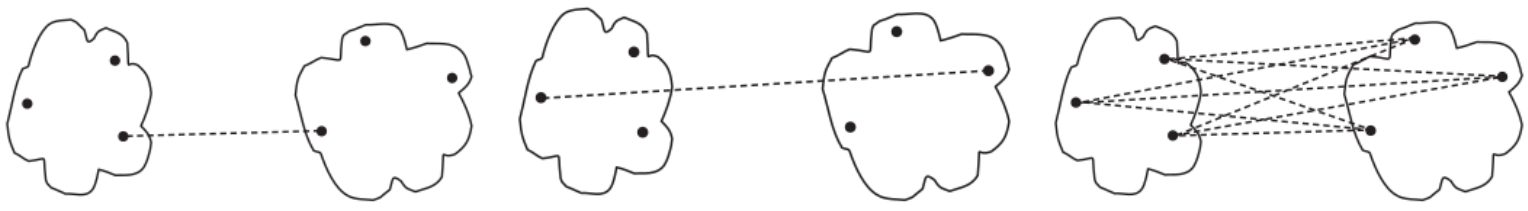
- Group average (genomsnitts länkning):

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

  där n_i och n_j är antalet observationer i kluster i och j.

- Wards/Centroid metod: Närheten defineras som hur mycket kvadrerade fel ökar när två kluster slås ihop.
  Samma kostnadsfunktion som i K-means.

Beräkning av avstånd mellan två kluster



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

Egenskaper

- Ingen global funktion att optimera.
- 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 senare.
- Närhetsmåttet kan påverka resultatet.
  - Extremvärden.
  - Brus.
- Passar bra för data som har en hierarkisk struktur.

Exempel

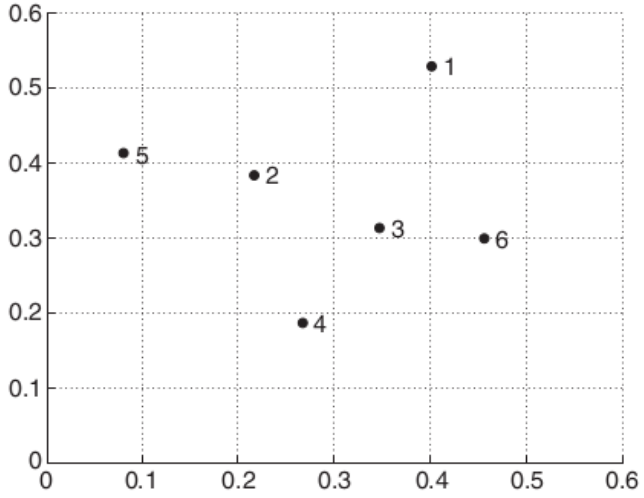


Figure 7.15. Set of six two-dimensional points.

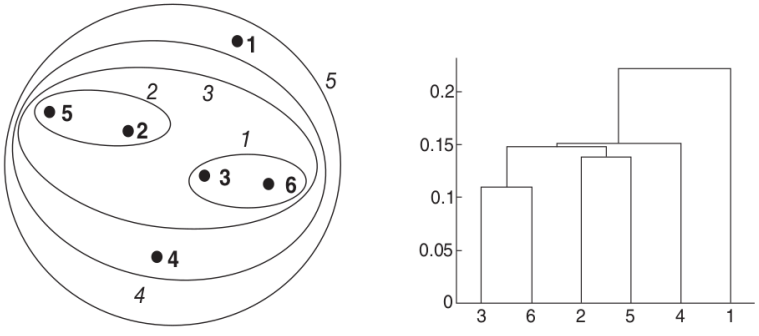
Point	x Coordinate	y Coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

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

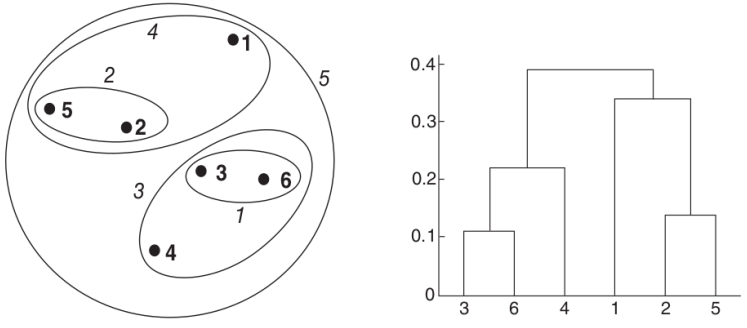
Table 7.4. Euclidean distance matrix for six points.

Exempel



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

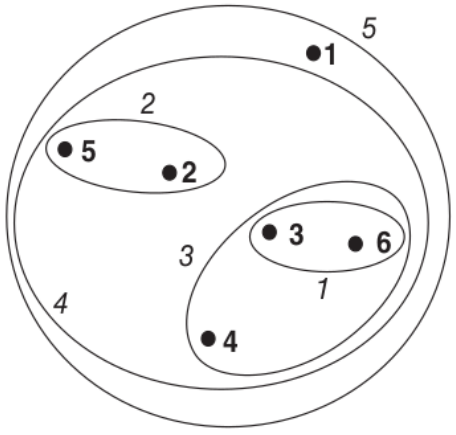
Figure 7.16. Single link clustering of the six points shown in Figure 7.15.



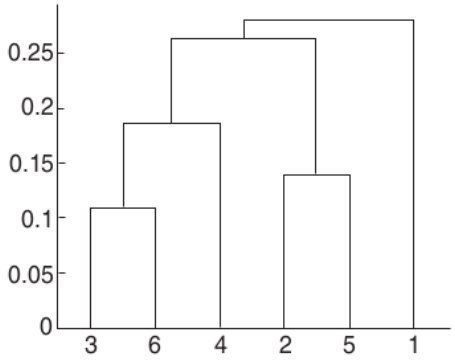
(a) Complete link clustering. (b) Complete link dendrogram.

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

Exempel

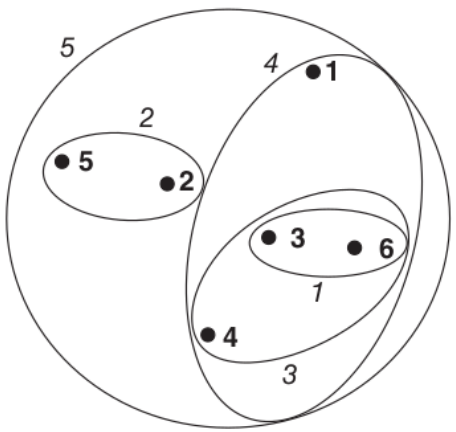


(a) Group average clustering.

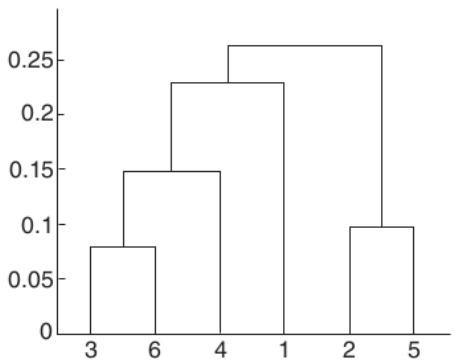


(b) Group average dendrogram.

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.