

# Klaszterezés, 2. rész

Csima Judit

BME, VIK,  
Számítástudományi és Információelméleti Tanszék

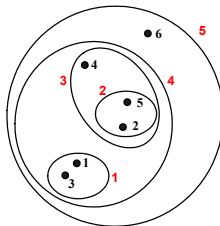
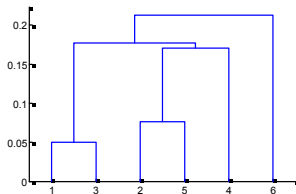
2014. április 17.

# Hierarchikus klaszterezés

- egymásba ágyazott klasztereket hoz létre
- nem kell előre megmondani, hogy hány klasztert szeretnénk, a teljes algo futása után a dendrogram vágásaival lehet előállítani különböző számú klasztereket
- két fő fajtája van: agglomerative és divisive

# Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits



# Két fajta hierarchikus klaszterzés

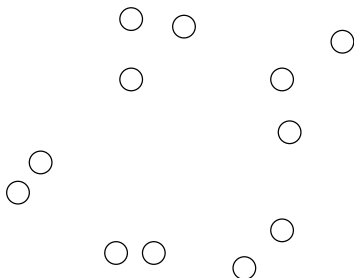
- agglomerative:
  - az elején minden pont külön klaszter
  - minden lépésben összevonom a két legközelebbi klasztert
- divisive:
  - az elején egy klaszter van az összes ponttal
  - minden lépésben valahogy szétvágom az egyik klasztert

# Agglomerative clustering

- kell egy távolságfogalom (vagy hasonlóság) a pontokra:  $L_2$  pl.
- azt is kell definiálnom, hogy mi lesz két klaszter távolsága: sok lehetőség, mindjárt nézzük őket, ezt tároljuk a proximity matrix-ban (itt is tárolhatok hasonlóságot vagy távolságot)
- algo:
  - az elején minden csúcs egy klaszter, a proximity matrix a pontok közti távolságot tartalmazza
  - amíg egynél több klaszter van: kiválasztom a két legközelebbi (leg hasonlóbb) klasztert, összevonom őket és frissítem a proximity matrix-ot

# Starting Situation

- Start with clusters of individual points and a proximity matrix



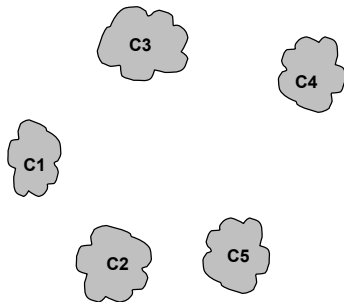
|    | p1 | p2 | p3 | p4 | p5 | ... |
|----|----|----|----|----|----|-----|
| p1 |    |    |    |    |    |     |
| p2 |    |    |    |    |    |     |
| p3 |    |    |    |    |    |     |
| p4 |    |    |    |    |    |     |
| p5 |    |    |    |    |    |     |
| .  |    |    |    |    |    |     |
| .  |    |    |    |    |    |     |
| .  |    |    |    |    |    |     |

**Proximity Matrix**



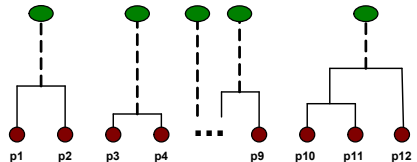
# Intermediate Situation

- After some merging steps, we have some clusters



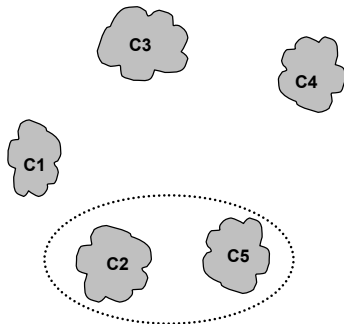
|    | C1 | C2 | C3 | C4 | C5 |
|----|----|----|----|----|----|
| C1 |    |    |    |    |    |
| C2 |    |    |    |    |    |
| C3 |    |    |    |    |    |
| C4 |    |    |    |    |    |
| C5 |    |    |    |    |    |

**Proximity Matrix**



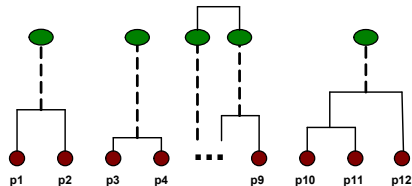
# Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



|    | C1 | C2 | C3 | C4 | C5 |
|----|----|----|----|----|----|
| C1 |    |    |    |    |    |
| C2 |    |    |    |    |    |
| C3 |    |    |    |    |    |
| C4 |    |    |    |    |    |
| C5 |    |    |    |    |    |

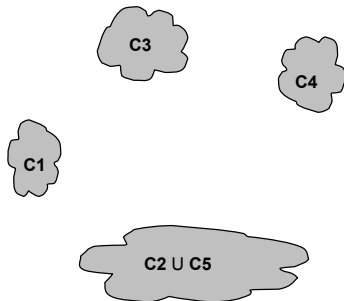
**Proximity Matrix**





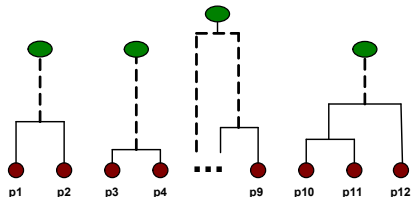
# After Merging

- The question is “How do we update the proximity matrix?”



|                |   |                        |           |           |
|----------------|---|------------------------|-----------|-----------|
|                |   | <b>C2<br/>U<br/>C5</b> | <b>C3</b> | <b>C4</b> |
| <b>C1</b>      |   | ?                      |           |           |
| <b>C2 U C5</b> | ? | ?                      | ?         | ?         |
| <b>C3</b>      |   | ?                      |           |           |
| <b>C4</b>      |   | ?                      |           |           |

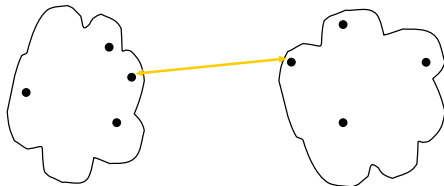
**Proximity Matrix**



# Hogyan definiáljuk két klaszter távolságát?

- MIN vagy single link:  
két klaszter távolsága/hasonlósága = a legkisebb távolság/legnagyobb hasonlóság, ami felvevődik két, külön klaszterben levő pont között
- MAX vagy complete link:  
két klaszter távolsága = a legnagyobb távolság/legkisebb hasonlóság, ami felvevődik két, külön klaszterben levő pont között

# How to Define Inter-Cluster Similarity

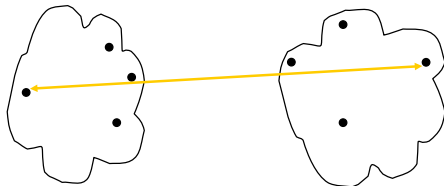


- ☐ MIN
- ☐ MAX
- ☐ Group Average
- ☐ Distance Between Centroids
- ☐ Other methods driven by an objective function
  - Ward's Method uses squared error

|    | p1 | p2 | p3 | p4 | p5 | ... |
|----|----|----|----|----|----|-----|
| p1 |    |    |    |    |    |     |
| p2 |    |    |    |    |    |     |
| p3 |    |    |    |    |    |     |
| p4 |    |    |    |    |    |     |
| p5 |    |    |    |    |    |     |
| .  |    |    |    |    |    |     |
| .  |    |    |    |    |    |     |
| .  |    |    |    |    |    |     |

**Proximity Matrix**

# How to Define Inter-Cluster Similarity



- ☐ MIN
- ☐ MAX
- ☐ Group Average
- ☐ Distance Between Centroids
- ☐ Other methods driven by an objective function
  - Ward's Method uses squared error

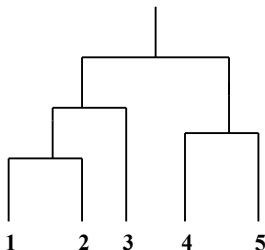
|    | p1 | p2 | p3 | p4 | p5 | ... |
|----|----|----|----|----|----|-----|
| p1 |    |    |    |    |    |     |
| p2 |    |    |    |    |    |     |
| p3 |    |    |    |    |    |     |
| p4 |    |    |    |    |    |     |
| p5 |    |    |    |    |    |     |
| .  |    |    |    |    |    |     |
| .  |    |    |    |    |    |     |
| .  |    |    |    |    |    |     |

**Proximity Matrix**

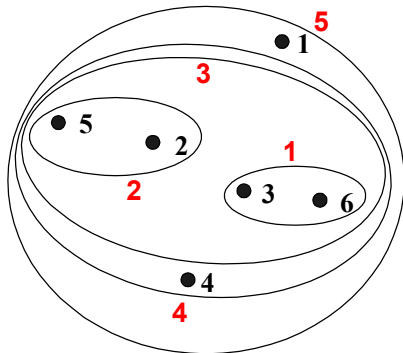
# Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points, i.e., by one link in the proximity graph.

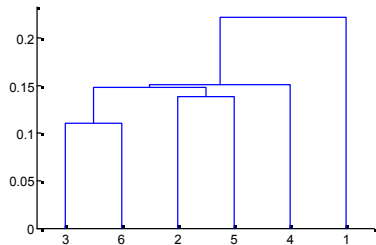
|    | I1   | I2   | I3   | I4   | I5   |
|----|------|------|------|------|------|
| I1 | 1.00 | 0.90 | 0.10 | 0.65 | 0.20 |
| I2 | 0.90 | 1.00 | 0.70 | 0.60 | 0.50 |
| I3 | 0.10 | 0.70 | 1.00 | 0.40 | 0.30 |
| I4 | 0.65 | 0.60 | 0.40 | 1.00 | 0.80 |
| I5 | 0.20 | 0.50 | 0.30 | 0.80 | 1.00 |



# Hierarchical Clustering: MIN



**Nested Clusters**

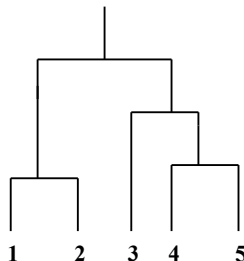


**Dendrogram**

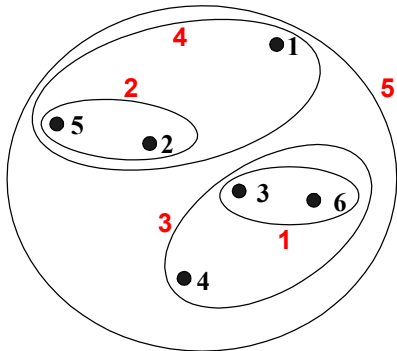
# Cluster Similarity: MAX or Complete Linkage

- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters

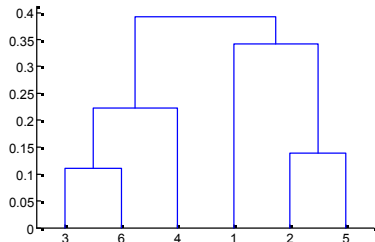
|    | l1   | l2   | l3   | l4   | l5   |
|----|------|------|------|------|------|
| l1 | 1.00 | 0.90 | 0.10 | 0.65 | 0.20 |
| l2 | 0.90 | 1.00 | 0.70 | 0.60 | 0.50 |
| l3 | 0.10 | 0.70 | 1.00 | 0.40 | 0.30 |
| l4 | 0.65 | 0.60 | 0.40 | 1.00 | 0.80 |
| l5 | 0.20 | 0.50 | 0.30 | 0.80 | 1.00 |



# Hierarchical Clustering: MAX



**Nested Clusters**



**Dendrogram**



# Klaszterek távolsága - átlagos távolság

- átlagos távolság: minden nem egy klaszterben levő pontpár távolságát figyelembe vesszük, ezt átlagoljuk

$$\bullet \text{ } dist(C_i, C_j) = \frac{\sum_{p \in C_i, q \in C_j} dist(p, q)}{|C_i||C_j|}$$

- ugyanez mehet hasonlósággal is távolság helyett

# Klaszterek távolsága centroidok használatával

- centroidok távolsága alapján összevonni: a legközelebbi centroidpárhoz tartozó klaszterek vonódnak össze
- Ward's method: azt a két klasztert vonjuk össze, amelyik esetén az SSE legkevesbé nő

# Hierarchikus klaszterezés: előnyök, hátrányok

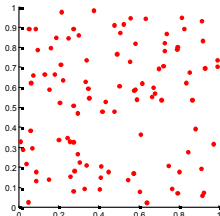
- ha két klasztert összevonok, azt már nem lehet visszacsinálni: érzékeny a zajra és outlier-ekre
- $n$  pont esetén  $O(n^3)$  a lépésszám, mert legfeljebb  $n$  darab  $O(n^2)$ -es menet van
- ha a klaszterezési feladat egy hierarchia keresést jelent (pl. taxonómia), akkor pont jó ez
- klaszterszám meghatározható az algo futása után
- sokszor Kmeans-szel együtt használják: egy kisebb mintán hierarchikus klaszterezés, az így kapott klaszterek centroidjaival kezdve pedig egy Kmeans ezután

# Klaszterezés értékelése, motiváció

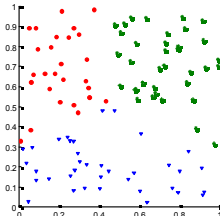
- szeretnénk valahogyan mérni, hogy mennyire jó egy klaszterzés
- miért?
  - el akarjuk kerülni, hogy ott is klasztert lássunk, ahol nincs: vannak-e valós klaszterek az adatban vagy csak mi találtunk?
  - két különböző klaszterezést ill. klaszterező algoritmust össze akarunk hasonlítani
- ez nehezebb, mint az osztályozás esetén volt

# Clusters found in Random Data

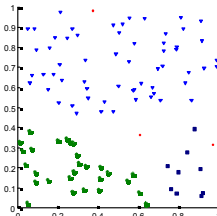
Random Points



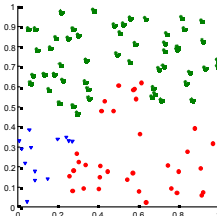
K-means



DBSCAN



Complete Link



- osztályozásnál volt jó mérőszám: ismertek a valódi címkék, ezek alapján accuracy, precision, recall, F-measure
- most (általában) külső segítség nélkül, csak az adatok alapján kell megítélni egy klaszterezés jóságát
- több megközelítés van

# Lehetséges értékelés külső címkéket használva

- van valami címkézés, ami ismert (szakértő is osztályozta az eseteket)
- ekkor hasonló van, mint az osztályozás: entrópia, purity, F-measure
- csak itt a címke a klaszter neve és az a kérdés, hogy ez mennyire esik egybe a szakértő csoportosításával

## External Measures of Cluster Validity: Entropy and Purity

Table 5.9. K-means Clustering Results for LA Document Data Set

| Cluster | Entertainment | Financial | Foreign | Metro | National | Sports | Entropy | Purity |
|---------|---------------|-----------|---------|-------|----------|--------|---------|--------|
| 1       | 3             | 5         | 40      | 506   | 96       | 27     | 1.2270  | 0.7474 |
| 2       | 4             | 7         | 280     | 29    | 39       | 2      | 1.1472  | 0.7756 |
| 3       | 1             | 1         | 1       | 7     | 4        | 671    | 0.1813  | 0.9796 |
| 4       | 10            | 162       | 3       | 119   | 73       | 2      | 1.7487  | 0.4390 |
| 5       | 331           | 22        | 5       | 70    | 13       | 23     | 1.3976  | 0.7134 |
| 6       | 5             | 358       | 12      | 212   | 48       | 13     | 1.5523  | 0.5525 |
| Total   | 354           | 555       | 341     | 943   | 273      | 738    | 1.1450  | 0.7203 |

**entropy** For each cluster, the class distribution of the data is calculated first, i.e., for cluster  $j$  we compute  $p_{ij}$ , the ‘probability’ that a member of cluster  $j$  belongs to class  $i$  as follows:  $p_{ij} = m_{ij}/m_j$ , where  $m_j$  is the number of values in cluster  $j$  and  $m_{ij}$  is the number of values of class  $i$  in cluster  $j$ . Then using this class distribution, the entropy of each cluster  $j$  is calculated using the standard formula  $e_j = \sum_{i=1}^L p_{ij} \log_2 p_{ij}$ , where the  $L$  is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e.,  $e = \sum_{i=1}^K \frac{m_j}{m} e_j$ , where  $m_j$  is the size of cluster  $j$ ,  $K$  is the number of clusters, and  $m$  is the total number of data points.

**purity** Using the terminology derived for entropy, the purity of cluster  $j$ , is given by  $\text{purity}_j = \max_i p_{ij}$  and the overall purity of a clustering by  $\text{purity} = \sum_{i=1}^K \frac{m_j}{m} \text{purity}_j$ .

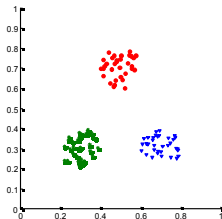


# Proximity matrix-szal vett korreláció

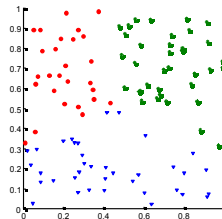
- proximity matrix: ki kihez van közel, két pont mennyire hasonló
- klasztermátrix (incidence matrix) :  $A_{ij} = 1$ , ha a két pont ugyanott van és 0, ha nem
- nézzük meg, hogy ez a két mátrix mennyire hasonló
- hogyan nézzük meg?
  - korreláció
  - a szemünkkel

# Measuring Cluster Validity Via Correlation

- Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets.



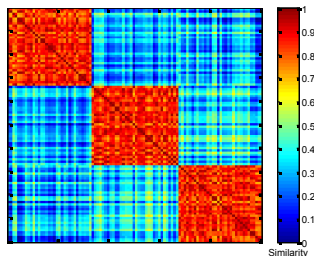
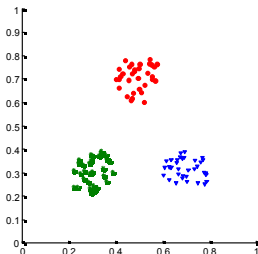
**Corr = -0.9235**



**Corr = -0.5810**

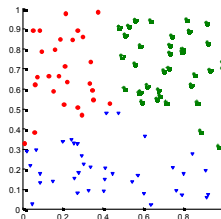
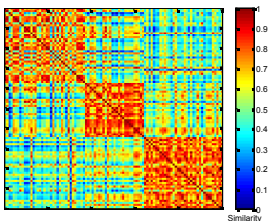
# Using Similarity Matrix for Cluster Validation

- Order the similarity matrix with respect to cluster labels and inspect visually.



# Using Similarity Matrix for Cluster Validation

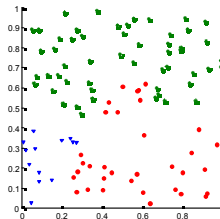
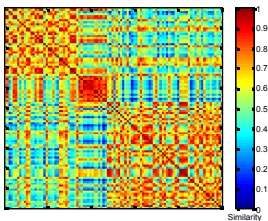
- Clusters in random data are not so crisp



**K-means**

# Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp



**Complete Link**

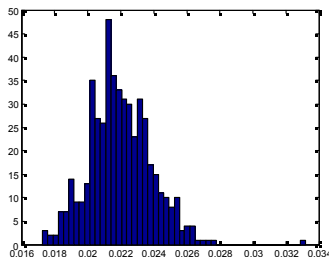
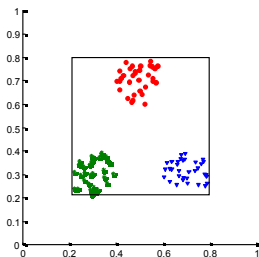
# SSE használata a klaszterezés értékelésére

- két klaszterezés összehasonlítható így: melyiknek kisebb az SSE-je?
- egy adott klaszterezés jóságának megítélése:
  - mekkora az esélye, hogy egy adott elemszámú random mintában ekkora SSE jön ki?
  - ehhez generálok sok véletlen mintát és megnézem, hogy milyen eloszlás lesz az SSE-re
  - ha úgy tűnik, hogy kicsi annak az esélye, hogy véletlenül olyan kicsi SSE jön ki, mint a miénk, akkor ez vleg egy jó klaszterezés

# Statistical Framework for SSE

## □ Example

- Compare SSE of 0.005 against three clusters in random data
- Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values



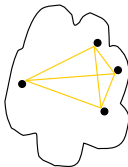
# Kohézió és szeparáltság mérése

- számszerűsítsük, hogy mennyire hasonlóak az egy klaszterbe eső pontok illetve mennyire különböznek a más klaszterbe esők
- kohézió mérésére pl. SSE vagy klaszteren belüli távolságok átlaga
- szeparációra pl. klaszterközi távolságok átlaga

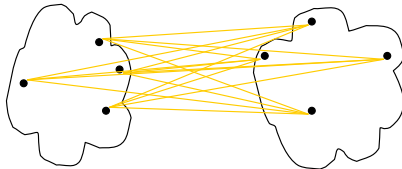


## Internal Measures: Cohesion and Separation

- A proximity graph based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion



separation