Kursus: Mitmemõõtmeline statistika

Seminar IX: Objektide grupeerimine

hierarhiline klasteranalüüs
k-keskmiste klasterdamine

Õppejõud: Katrin Niglas
PhD, dotsent
informaatika instituut

TALLINNA ÜLIKOOL

Objektide grupeerimine

Eesmärk (ehk miks **objekte koondada**):

- Omavahel sarnaste objektide leidmine
  (sarnased = valitud tunnuste lõikes ligilähedaselt samu vastuseid/väärtusi omavad vastajad/objektid)
- Tüpoloogiate loomine

TALLINNA ÜLIKOOL

Objektide grupeerimine

Võimalikud meetodid:

Vähe objekte:

- Hierarhiline klasteranalüüs
- k-keskmiste klasteranalüüs

Palju objekte:

- k-keskmiste klasteranalüüs

Kategoriaalsed tunnused:

- Tunnuste poolt defineeritud alamgruppidesse jagamine (abiks mitmemõõtmelised sagedustabelid)

TALLINNA ÜLIKOOL

Hierarhiline klasteranalüüs

Hierarhilise klasteranalüüsi idee seisneb selles, et samm-sammult hakatakse kokku panema neid objekte, mis on omavahel kõige sarnasemad st millel on valitud tunnuste lõikes kõige sarnasemad väärtused.

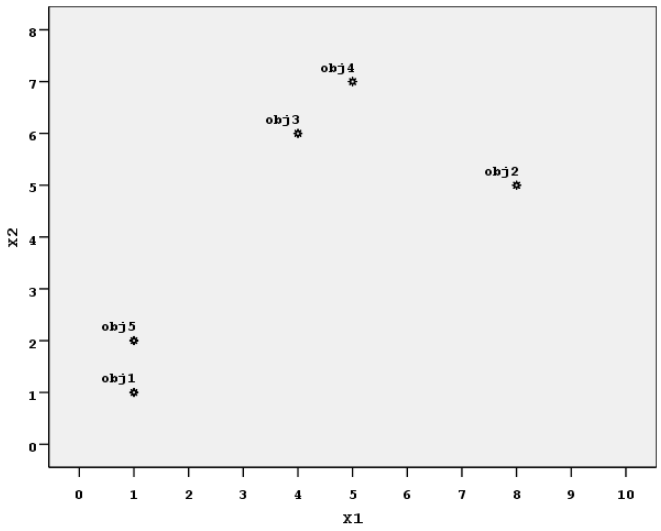
=> Esimesel sammul iga objekt omaette – viimasel sammul kõik koos st üks klaster/grupp – põhiküsimus: millal kokkupanemine lõpetada!?

Matemaatiliselt: Vaja valida objektide sarnasuse/kauguse mõõt ja klastrite vahelise kauguse arvutamise loogika

NB! Kui valitud tunnused on mõõdetud erinevatel skaaladel, siis on tihti otstarbekas tunnused standardiseerida, et vältida nende skaalade domineerimist kauguse leidmisel, mille ühikud ja absoluutne hajuvus on suuremad.

TALLINNA ÜLIKOOL

Hierarhiline klasteranalüüs



TALLINNA ÜLIK

Hierarhiline klasteranalüüs

Kauguse/sarnasuse mõõduks *Measure/Interval:*

$Obj_1(x_{11};x_{21};...;x_{p1})$ $Obj_2(x_{12};x_{22};...;x_{p2})$

- (Squard) Euclidean distance – Eukleidiline kaugus

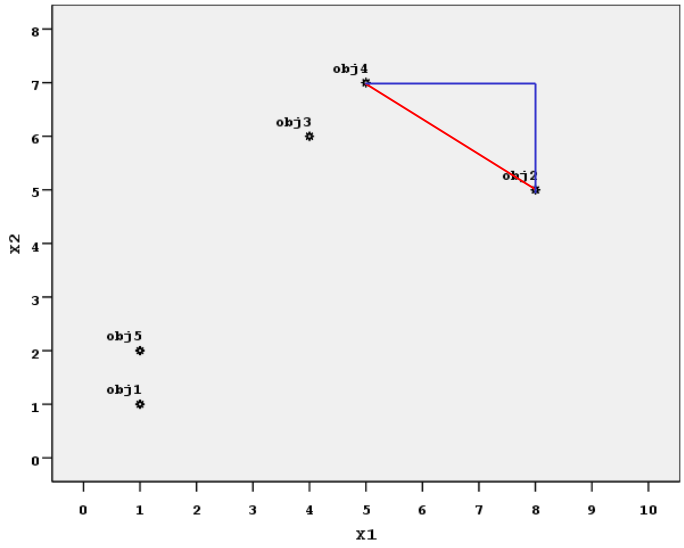
$$D_{Eukleidiline} = \sqrt{(x_{11}-x_{12})^2+(x_{21}-x_{22})^2+...+(x_{p1}-x_{p2})^2}$$

- Minkowski

$$D_{Minkowski} = \left[\sum_{i=1}^{p}\left|x_{i1}-x_{i2}\right|^m\right]^{\frac{1}{m}} \qquad D_{City-block} = \sum_{i=1}^{p}\left|x_{i1}-x_{i2}\right|$$

TALLINNA ÜLIKOOL

Hierarhiline klasteranalüüs



TALLINNA ÜL

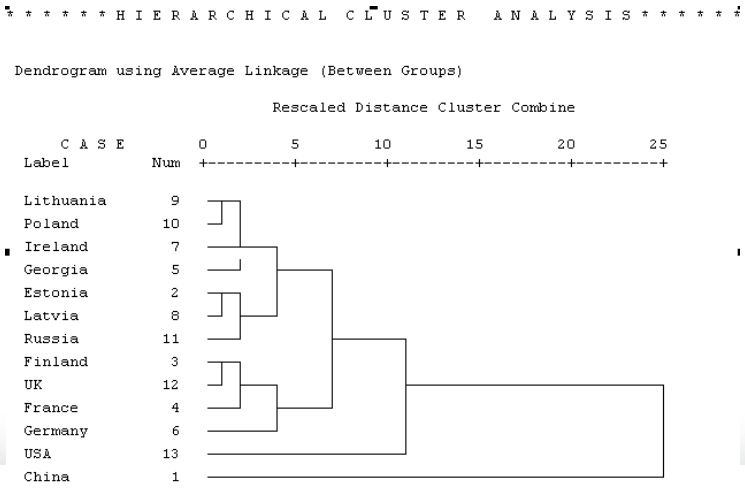
Hierarhiline klasteranalüüs

Klastrite ühendamise meetod väljal *Cluster Method:*

- *Between-groups linkage* klastrite vaheline kaugus arvutatakse üksiktunnuste kauguste/sarnasuste keskväärtusena
- *Nearest neighbor* klastrite vaheline kaugus arvutatakse erinevates klastrites olevate omavahel kõige sarnasemate objektide vahelise kaugusena
- *Centroid clustering* klastrite vaheline kaugus arvutatakse klastri raskuskeskmete vahel
- *Ward's method* teistest erinev loogika (ANOVA'le sarnane)
  minimiseeritakse klastrite sisest hajuvust
  E= summarne objektide kaugus klastri keskpunktist
  algul kõik eraldi ja E=0
  igal sammul pannakse kokku need objektid, mille puhul E kasvab kõige vähem

TALLINNA ÜLIKOOL

Hierarhiline klasteranalüüs - dendrogramm



TALLINNA ÜLIKOOL

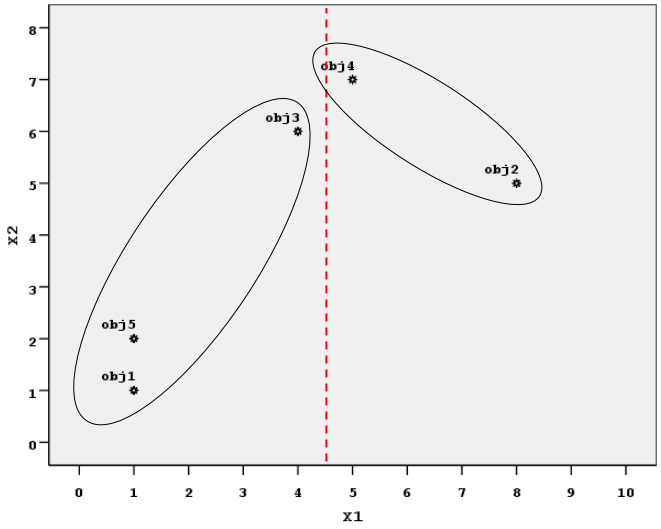
k-keskmiste klasterdamine

k-keskmiste klasterdamise algoritm on samuti lihtne:

- määrata klastrite arv
- jagada objektid esialgsetesse klastritesse
- arvutada välja klastrite keskpunktid
- võrrelda igat objekti kõigi klastrite keskpunktidega;
  kui osutub, et objekti kaugus mõne muu klastri keskpunktist on väiksem kui selle klastri keskpunktist, milles ta parasjagu asub, siis tuleb objekt teise klastrisse ümber tõsta
- peale objekti ümbertõstmist tuleb klastrite keskpunktid uuesti arvutada ja jätkata protsessi niikaua kui kõik objektid on klastris, mille keskpunktile nad kõige lähemal asuvad.

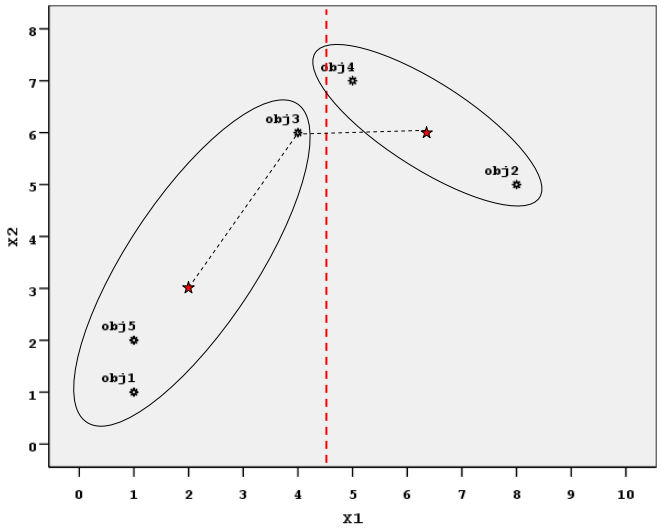
TALLINNA ÜLIKOOL

k-keskmiste klasterdamine



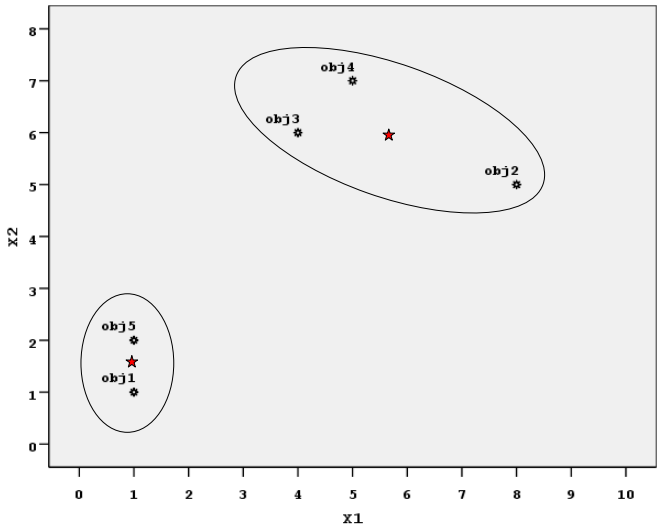
TALLINNA ÜLIK

k-keskmiste klasterdamine



TALLINNA ÜLIK

k-keskmiste klasterdamine



TALLINNA ÜLIK

Klasteranalüüs – pane tähele!

- Pole ühte ainuõiget lahendit => tõlgendus + stat.näitajad
- Kui andmetes ei ole sisulist grupeeritust, siis võib jagamine olla "meelevaldne" ja sõltuda (suurel määral) klasterdamise meetodist
- Tulemus võib sõltuda objektide sorteerimise järjekorrast
- Ekstreemsed väärtused => ühe-kahe objektiga klastrid
- Erinevad skaalad => standardiseerida
- Eksploratiivne meetod => F-testi e ANOVA't ei tohi klasteranalüüsi juures tõlgendada kui olulisustesti
- Kaalu, kas klasterdamise aluseks olevad tunnused on valitud "mõistlikult" (sobivad sisuliselt kokku, ükski tähtis tunnus ei puudu mudelist, jne)

TALLINNA ÜLIKOOL

Kursus: Mitmemõõtmeline statistika

Seminar X: Tunnuste ja objektide grupeerimine
Klasteranalüüs
Faktoranalüüs

Õppejõud: Katrin Niglas
PhD, dotsent
informaatika instituut

TALLINNA ÜLIKOOL

Praktikum: tunnuste ja objektide grupeerimine

Andmestik:	Muulased.sav
Raamat:	"Vene küsimus ja Eesti valikud", toim. Mati Heidmets, TPÜ Kirjastus 1998
Artikkel:	"Usaldus ja usaldamatus rahvussuhetes", Jüri Kruusvall, lk 29-76
Faktoranalüüs:	peatükk 2.3. Eesti elanike etniline häiritus (lk 36)
Klasteranalüüs:	peatükk 2.4. Eesti elanike etnilise häirituse tüpoloogia (lk41)

NB! Tabelid on artikli lõpus!

TALLINNA ÜLIKOOL

Praktikum: tunnuste ja objektide grupeerimine

Teine näide faktor- ja klasteranalüüsi kasutanud uurimuse põhjal kirjutatud artiklitest:

Ehala, M., Niglas, K. (2004) Eesti koolinoorte keelehoiakud. *Akadeemia*, 2004, 10, lk 2115-2143.

Ehala, Martin; Niglas, Katrin (2006). Language attitudes of Estonian secondary school students. *Journal of Language, Identity and Education*, 5(3), 209 - 227.

TALLINNA ÜLIKOOL