

Klyngeanalyse

ISTx1003 Statistisk læring og Data Science

Stefanie Muff, Institutt for matematiske fag

November 12, 2021

Plan for i dag

- Hva er klyngeanalyse
- Læringsmål, pensum og læringsressurser
- Avstandsmål
- K-gjennomsnitt (“K-means”) klyngeanalyse
- Bruk av klyngeanalyse på et bilde (prosjektet fra i fjor)
- Hierarkisk klyngeanalyse
- Informasjon om prosjektet

Eksempel 1: Genaktivitet

- $n = 81$ celleprøver fra kreftsvulster til ulike pasienter
- genaktivitet for $p = 12957$ gener

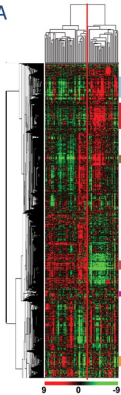
Spørsmål:

Hvilke celleprøver fra brystkreftpasienter ligner hverandre mest?

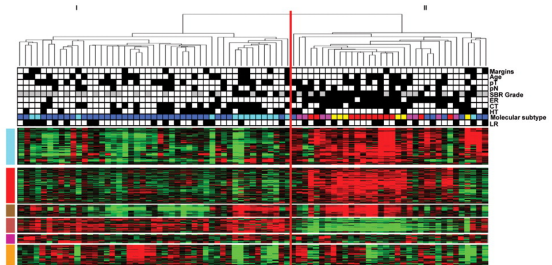
Kan vi finne ukjente klynger (av celleprøver) i dataene?

Dette kan hjelpe for å forutse sannsynligheten for en tilbakefall.

A



B



$$X = p \times n = \text{gener} \times \text{prøver} .$$

Finn ut mer: <https://cgp.iiajournals.org/content/8/4/199>

Eksempel 2: Proteininteraksjonsnettverk

Kan vi finne klynger med relatert funksjon?

MUFF, RAO, AND CAFLISCH

PHYSICAL REVIEW E 72, 056107 (2005)

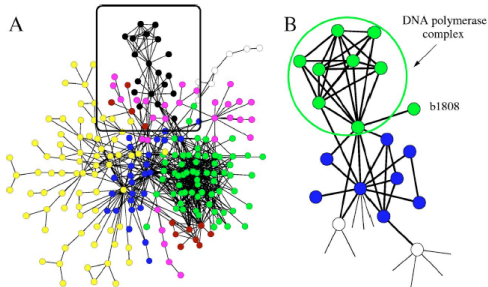


FIG. 4. (Color online) (a) Largest connected component of the PPI of *E. coli*. The colors represent the clusterization found by optimizing modularity. (b) LQ clusterization of the black Q cluster. The green circle contains proteins belonging to the DNA polymerase complex. The unknown protein b1808 is assigned to this complex according to LQ while the complete Q cluster is heterogeneous.

Eksempel 3: Bildanalyse

Det var en prosjektoppgave i fjor.

Mål:

Å bruke klyngeanalyse til å fjerne detaljer og støy - ved å dele pikslene inn i to eller flere klynger.

Todo: Ev adapt with task from this year.

Læringsmål

- forstå hvorfor det er interessant å gjøre klyngeanalyse
- kjenne igjen situasjoner der klyngeanalyse vil være en aktuell metode å bruke
- kjenne begrepene avstandsmål, koblingstype, dendrogram
- forstå algoritmen for å utføre K-gjennomsnitt-klyngeanalyse og hierarkisk klyngeanalyse
- forstå hvordan klyngeanalyse utføres i Python
- kunne besvare oppgave 3 av prosjektoppgaven på en god måte!

Læringsressurser

Tema Klyngeanalyse:

- **Kompendium:** Klyngeanalyse (pdf og html, by Mette Langaas)
- **Korte videoer:** (by Mette Langaas)
 - Klyngeanalyse (8:43 min)
 - Hierarkisk klyngeanalyse (11:26 min)
 - K-gjennomsnitt-klyngeanalyse (8:38 min)
- Denne forelesningen
- **Disse slides** med notater

Som alltid se her:

<https://wiki.math.ntnu.no/istx1003/2021h/start>

Klyngeanalyse – hva er det?

Vi har data

$$X : n \times p$$

men *ikke* noen respons Y . *Ikke-veiledet = unsupervised*

Mål:

- Finn ukjente klynger i dataene.
- Observasjoner innen hver klynge er mer lik hverandre enn observasjoner fra ulike klynger.

Hva skal vi bruke resultatene fra klyngeanalysen til?

- Bildet: Fjerne støy eller, spare lagringsplass
- Medisin: Finne subgrupper av en sykdom → relevant for behandling?

Klyngeanalyse – hva er det?

Generelt: Finne *struktur* i dataene.

Kan vi stole på resultatene? Hvor robuste er de?

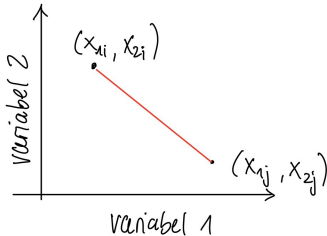
→ Fortsatt et forskningsområde!

Avstandsmål

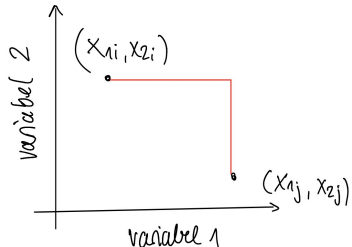
Før en klyngeanalyse må vi først definere en *avstand* mellom to datapoeng.

To populære avstandsmål:

Euklidsk



City-block (=Manhattan)



Euklidsk

$$D_E(i, i') = \sqrt{\sum_{j=1}^p (x_{ji} - x_{ji'})^2}$$

City-block (=Manhattan)

$$D_M(i, i') = \sum_{j=1}^p |x_{ji} - x_{ji'}|$$

Avstandsmål i mer enn 2 dimensjoner: Enkelt å regne, men litt vanskelig å forestille seg.

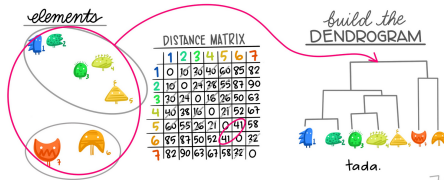
Metoder for klyngeanalyse

Det fins ganske mange metoder, men vi ser på to som er (mest?) populær:

K-gjennomsnitt klyngeanalyse

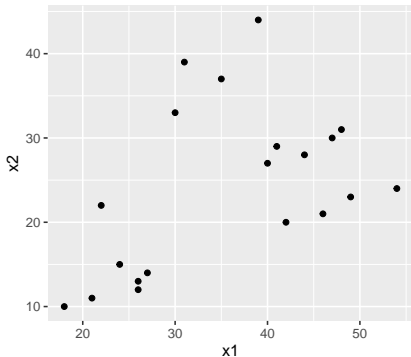


Hierarkisk klyngeanalyse



K-gjennomsnitt klyngeanalyse

- Finn K ukjente klynger i dataene.



- Alle observasjoner skal være medlem i akkurat en klynge
- Variasjonen innen hver klynge skal være så liten som mulig

Variasjon innen en klynge k

- K klynger $C_1, \dots, C_k, \dots, C_K$.
- Antall observasjoner i klynge k : $|C_k|$.
- Variasjon in klynge k :

$$\frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

Optimeringsproblem

Vi vil *minimere* variasjon over *alle klynger*:

$$\sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

Nyttig sammenhang som er grunnlag for k -gjennomsnitt algoritme

$$\begin{aligned} & \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \\ &= \sum_{k=1}^K 2 \sum_{i \in C_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2, \end{aligned}$$

med *klyngecentroide* i klynge k : $\bar{x}_k = (\bar{x}_{k1}, \dots, \bar{x}_{kp})$

K-gjennomsnitt algoritme

- Start med å velge antall klynker K .
- Tilordne hver observasjon til en klynge
 - Mange muligheter
 - tilfeldig velge ut K observasjoner og sette disse som klyngesentroider
 - tilfeldig klynger
 - og deretter tilordne der resterende observasjoner til klyngen med nærmeste klyngesentroide.
- **Repeter** (iterativt) *til ingen observasjoner endrer klyngemedlemskap*:
 1. For hver klynge regn ut klyngesentroiden
 2. Tilordne hver observasjon til klyngen til nærmeste klyngesentroide

Illustrasjon av K -gjennomsnitt algoritme ($K = 3$)

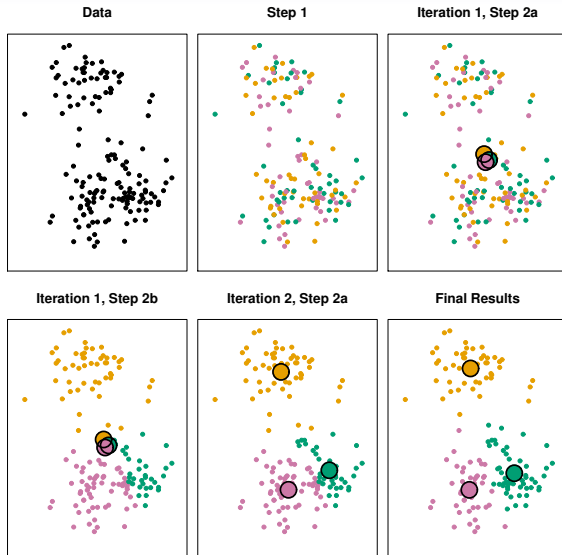


Fig. 10.6 fra “An Introduction to Statistical Learning with Applications in R”, James et al 2013.

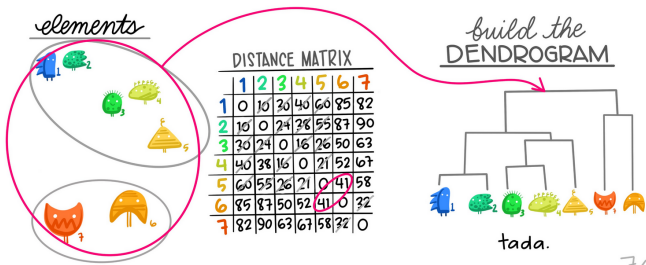
K-gjennomsnitt-algoritmen i Python

Todo

Prosjektoppgaven

Todo

Hierarkisk klyngeanalyse

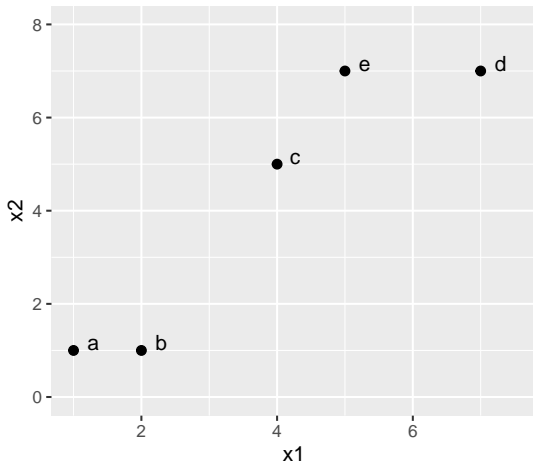


Artwork by @allison_horst

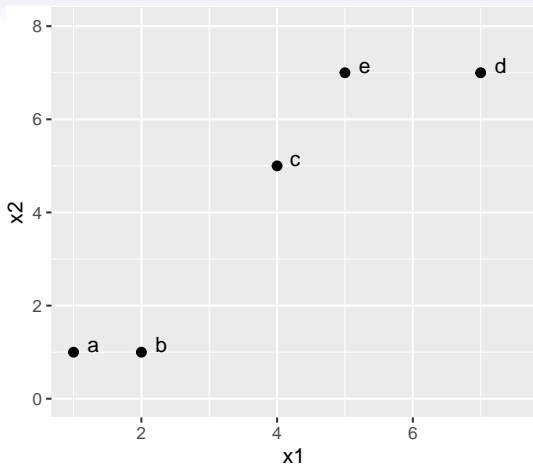
Eksempel

$$n = 5, p = 2$$

x1	x2	name
1	1	a
2	1	b
4	5	c
7	7	d
5	7	e

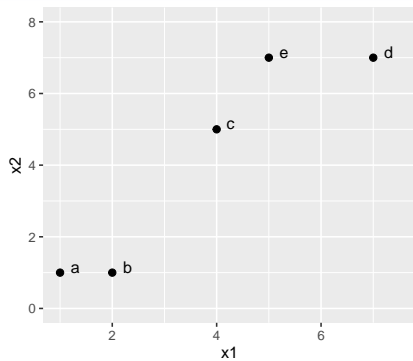


x1	x2	name
1	1	a
2	1	b
4	5	c
7	7	d
5	7	e



- 1) Velg avstandsmål
- 2) Regn ut avstanden mellom alle par av observasjoner
- 3) Plasser avstandene inn i en $n \times n$ matrise

Avstandsmatrise (Euklidsk avstand)



name	a	b	c	d	e
a	0.0	1.0	5.0	8.5	7.2
b	1.0	0.0	4.5	7.8	6.7
c	5.0	4.5	0.0	3.6	2.2
d	8.5	7.8	3.6	0.0	2.0
e	7.2	6.7	2.2	2.0	0.0