

Klyngeanalyse

ISTx1003 Statistisk læring og Data Science

Sara Martino, Institutt for matematiske fag

Oktober 29, 2024

Takk til

Mette Langaas (pensumvideoer, kompendium, slides og inspirasjon)
Stefanie Muff (slides og inspirasjon) Ingeborg Hem Sørmoen (slides og inspirasjon)

Mine slides bygger på materiale fra både Stefanie, Mette og Ingeborg, som underviste dette kurset i 2021-2023

Praktisk om denne zoom-forelesningen

- Zoom-webinar (som i fellesdelen)
- Forelesningen blir tatt opp og lagt ut
- Dere kan bruke chatten for å sende spørsmål til meg, jeg skal prøve å følge med så godt jeg kan.
- Jeg oppfordrer dere til å sitte sammen med andre studenter og se forelesningen!
- Alle info ligger i wiki sida:
<https://wiki.math.ntnu.no/istx100y/2024h/1003>

Plan - Forelesninger

- **Mandag 21. oktober 15:15-16:00:** Introduksjon
- **Tirsdag 22. oktober 14:15-16:00:** Oppgave 1 (multippel lineær regresjon)
- **Mandag 28. oktober 15:15-16:00** Oppgave 2 (klassifikasjon) og oppgave 3 (klyngeanalyse)
- **Tirsdag 29. oktober 14:15-16:00** Oppgave 3 fors. (?), Q&A og praktisk informasjon
- Alle de 4 forelesningene vil være hel-digitale
- Alle forelesningene tas opp, tentative slides legges ut før forelesningen og slides med notater legges ut etter forelesningen
- Forelesningene vil i all hovedsak utdype prosjektoppgavene. Pensum dekkes av kompendier og tilhørende videoer, og blir ikke repetert her.

Plan for i dag (tema “Klyngeanalyse”)

- 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

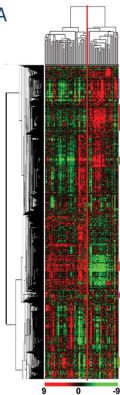
Spørsmål:

Hvilke celleprøver fra brystkreftpasienter ligner hverandre mest?

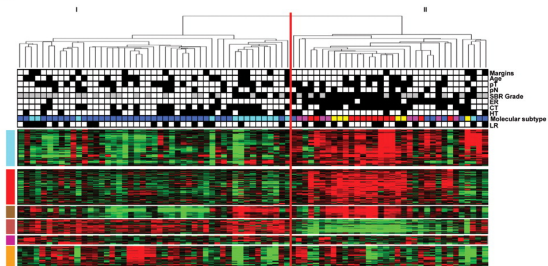
Kan vi finne ukjente klynger (av celleprøver) i dataene?

Dette kan hjelpe for å forutsi 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>

->

->

Læringsmål

Etter du har lest kompendiet, sett videoene som er laget og deltatt på zoom-forelesning skal du:

- kunne 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 hierarkisk klyngeanalyse og k -gjennomsnitt-klyngeanalyse
- kunne utføre klyngeanalyse i Python
- **kunne svare godt på oppgave 3 i den tellende prosjektoppgaven**

Læringsressurser

Tema Klyngeanalyse:

- Kompendier (laget av Mette Langaas)
- Tilhørende videoer (laget av Mette Langaas)
- Zoom-forelesningene (med slides)

Kompendier og tilhørende videoer ligger allerede på *Wiki sida*.

Zoom-forelesningene og slides legges også ut. Mange andre tilgjengelige kilder på nett også, men jeg garanterer ikke at de dekker de riktige delene av temaene.

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 $\rightarrow$ relevant for behandling?

Klyngeanalyse – hva er det?

Generelt: Finne *struktur* i dataene.

Kan vi stole på resultatene? Hvor robuste er de?

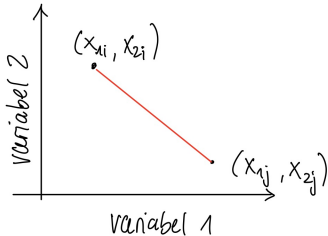
→ Fortsatt et forskningsområde!

Avstandsmål

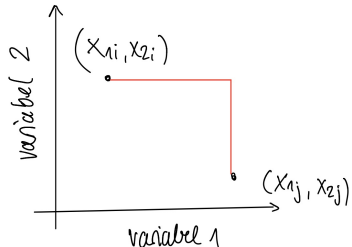
Før en klyngeanalyse må vi først definere en *avstand* mellom to datapoeng.

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.

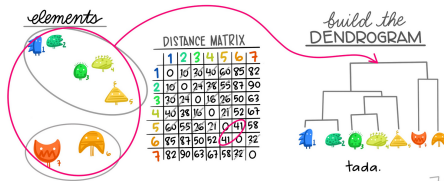
Metoder for klyngeanalyse

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

K-gjennomsnitt klyngeanalyse

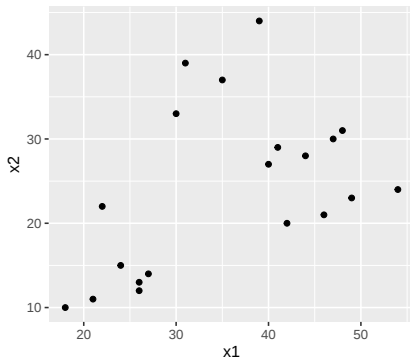


Hierarkisk klyngeanalyse



K-gjennomsnitt klyngeanalyse

- Finn K ukjente klynger i dataene.



- Alle observasjoner skal være medlem i akkurat *én* 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*:

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

K-gjennomsnitt algoritme

- Start med å velge antall klynker K .
- Tilordne hver observasjon til en klynge
 - Mange muligheter
 - å tilfeldig velge ut K observasjoner og sette disse som klyngesentroider, og deretter tilordne de resterende observasjonene til klyngen med nærmeste klyngesentroide.
 - tilfeldig klynger
- **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 ($K = 3$)

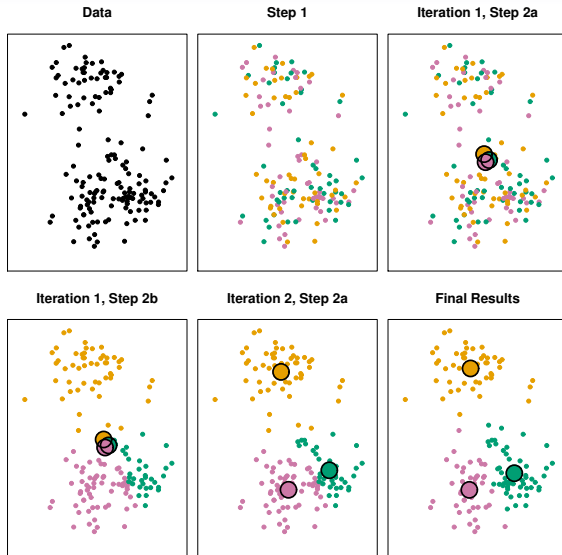
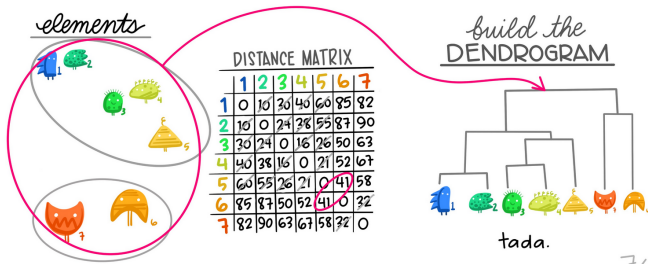


Fig. 10.6 fra “An Introduction to Statistical Learning with Applications in R”, James et al 2013.

Hierarkisk klyngeanalyse

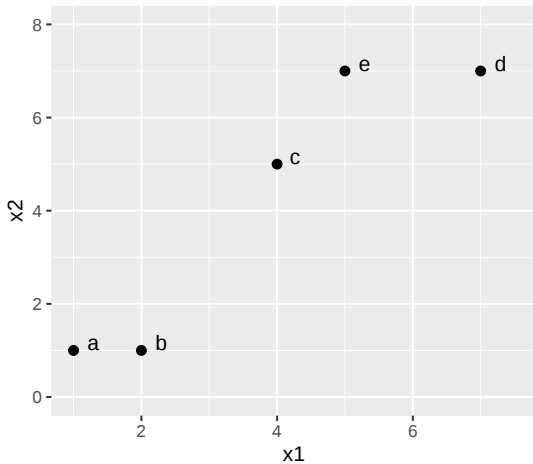


Artwork by @allison_horst

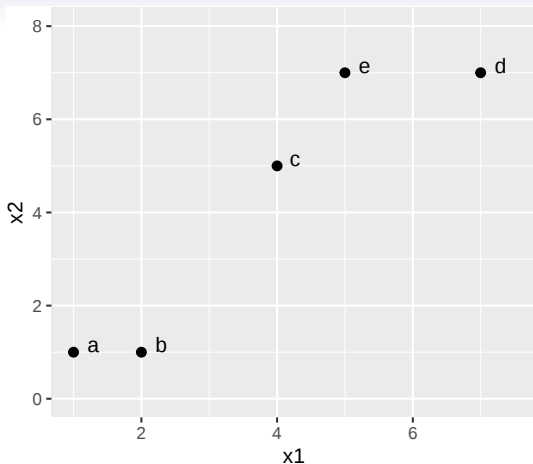
Eksempel

$$n = 5, p = 2$$

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

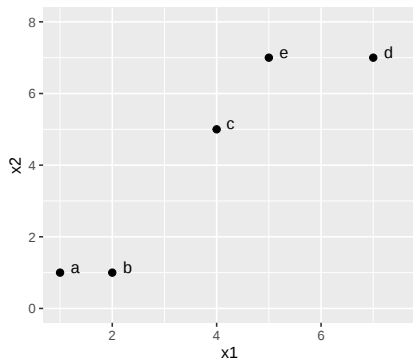


x1	x2	name
1	1	a
2	1	b
4	5	c
7	7	d
5	7	e



- 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)

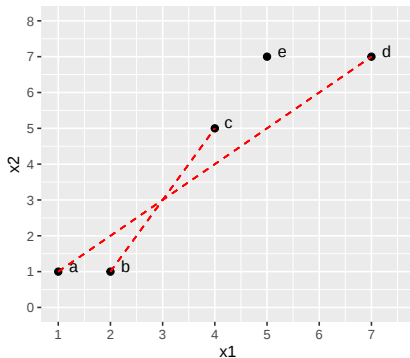


name	a	b	c	d	e
a	0.0	1.0	5.0	8.5	7.2
b	1.0	0.0	4.5	7.8	6.7
c	5.0	4.5	0.0	3.6	2.2
d	8.5	7.8	3.6	0.0	2.0
e	7.2	6.7	2.2	2.0	0.0

Avstand mellom klynger?

Tre populære typer avstandsmål:

- **Singel kobling:** minimal avstand
- **Komplett kobling:** maksimal avstand
- **Gjennomsnittskobling:** gjennomsnittlig avstand



Algoritme for hierarkisk klyngeanalyse - I

Før algoritmen starter man bestemme seg for

- hvilket avstandsmål skal brukes (f.eks: euklidsk, city block, korrelasjon,...)
- hvilken koblingstype skal brukes (f.eks: singel, komplett, gjennomsnitt, sentroide,...)

og regne ut avstandsmatrisen mellom alle observasjoner.

Algoritme for hierarkisk klyngeanalyse – II

Behandle hver observasjon som om den var sin egen klynge (det er da n klynger).

1. Slå sammen

Finn de to klyngene som er nærmest hverandre og slå dem sammen til en klynge.

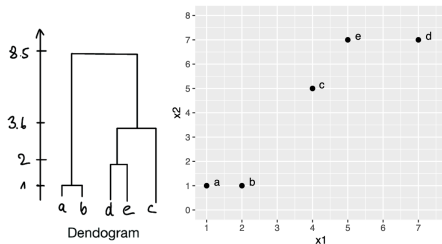
2. Beregn avstander

Beregn nye parvise avstander mellom alle klynger ved bruk av valgt avstandsmål og koblingstype.

Repeterer til alle observasjonene er i samme klynge.

Resultat fra vår eksempel (dendrogram)

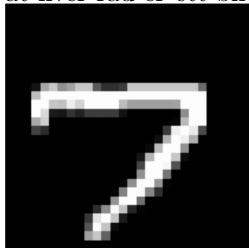
Ved bruk av komplett kobling (maksimal avstand):



Oppgave 3

Hver observasjon er et bilde av et tall, med $20 \times 28 = 784$ piksler, så vi regner ut avstand i 784 dimensjoner (i motsetning til 2 dimensjoner på et ark, eller 3 i rommet).

Når vi lagrer observasjonene, bruker vi én kolonne for hver piksel, slik at hver rad er ett bilde i datamatriksen.



Oppgave 3

OBS! Klyngesentroiden har samme dimensjon som dataene!

Hvis dataene er 2-dimensjonale punkter, er sentroiden også et 2-dimensjonalt punkt.

Hvis dataene er 784-dimensjonale tall, er sentroiden også 784-dimensjonal. Den kan visualiseres på samme måte som dataene.

Oppgave 3

10 korte, konkrete spørsmål. Omtrent all kode dere trenger står i en Jupyter notatbok.

Bruk svarark (enklere for både dere og oss).

Teller 20% av prosjektet, slik at hvert spørsmål teller 2% hver.

Innlevering av projektet

- Dere skal lever **to pdf-filer**. En for oppgave 1 og en for oppgave 2 og 3.
- Begge filene skal være pdf-filer, og dere er selv ansvarling for at alle svar er med i de to filene. Filene skal være ryddig og oversiktige.

**Frist for innlevering er 18 november 12:00 i
Inspira**

Videre de neste to ukene

- Hvis dere ikke har gjort det: Se på de korte videoene for hvert tema.
- Jobb med prosjektoppgaven. Husk at frist for innlevering av prosjektet til Inspira er **18 november 12:00 i Inspira**.
- Bruk veiledning timer (see <https://wiki.math.ntnu.no/istx100y/2024h/1003>)
- Husk også mattelab forumet – men bare for korte spørsmål. Lange spørsmål fungerer best med direkt interaksjon.