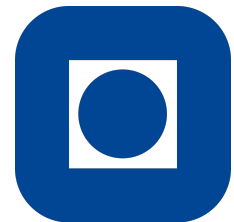


Version fra zoom-forelesningen
11.11.2020

ISTx1003: Statistisk læring og data science

Zoom-forelesning 4: Klyngeanalyse

Mette Langaas, IMF/NTNU. 11.11.2020



NTNU

Plan: Klyngeanalyse

1. Hva er klyngeanalyse?
2. Læringsmål, pensum og læringsressurser
3. Avstandsmål
4. K-gjennomsnitt klyngeanalyse
5. Bruk av klyngeanalyse på et bilde (prosjektet)
6. Hierarkisk klyngeanalyse
7. Har vi dekket pensum?
8. Innspurt prosjektmodulen mot 29.11 kl 23.59

(Inkludert 15 minutters pause)

Eksempel: genaktivitet

- $n=64$ celleprøver fra kreftsvulster til ulike pasienter
- genaktivitet for $p=6830$ gener



Hvilke celleprøver ligner hverandre mest? *to brystkreftprøver?*

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

Eksempel: genaktivitet

100 av
de 6830
gener

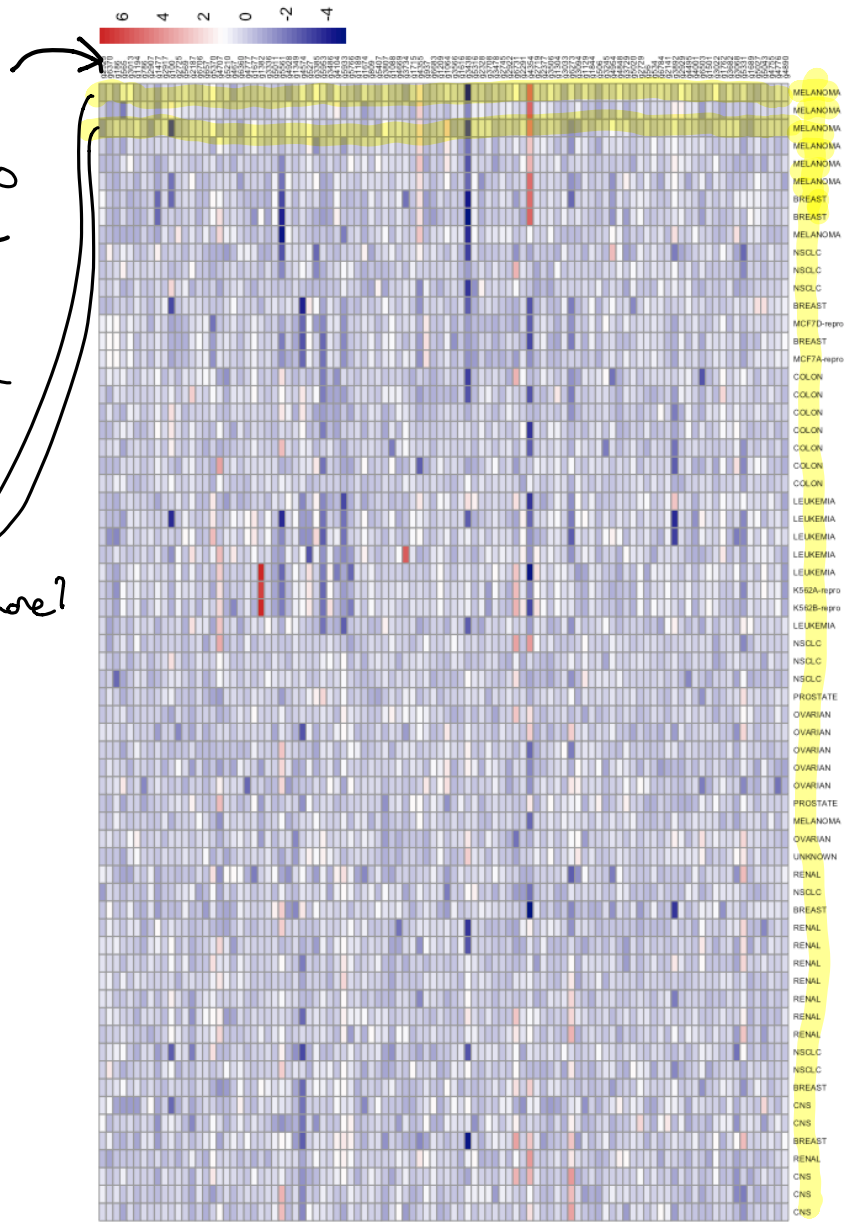
rodt = høyt regulert
blått = lavt regulert

er disse
to like værende?

64 prøver

$$X : n \times p$$

6830
variabler



Eksempel: bildeanalyse

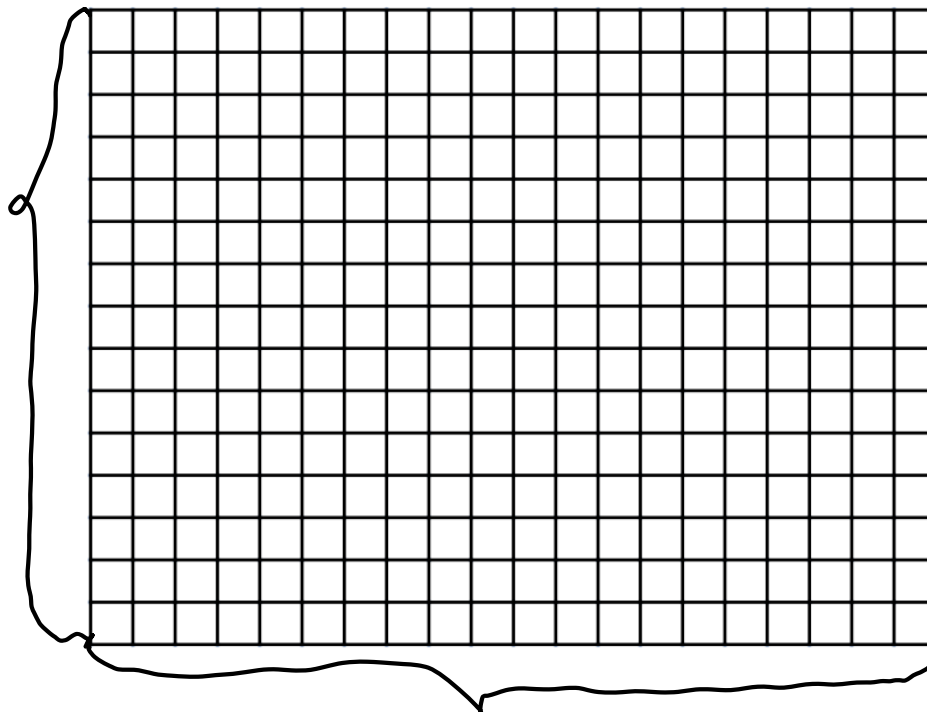
Kan vi bruke klyngeanalyse til å fjerne detaljer og støy - ved å dele pikslene inn i to eller flere klynger?



Eksempel: bildeanalyse

'china.jpg'

427



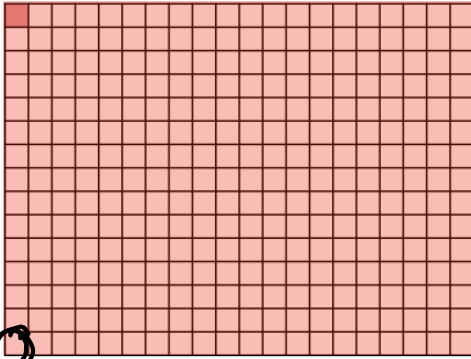
640

"tabell, matrise"

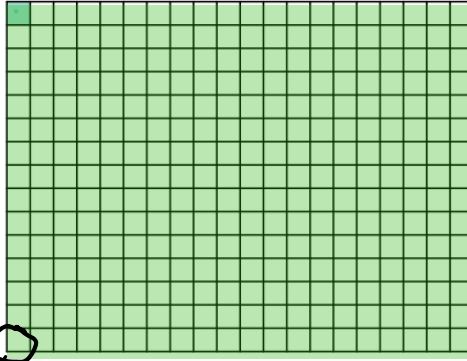
$$427 \cdot 640 \approx \frac{1}{4} \text{ mill}$$

273 280

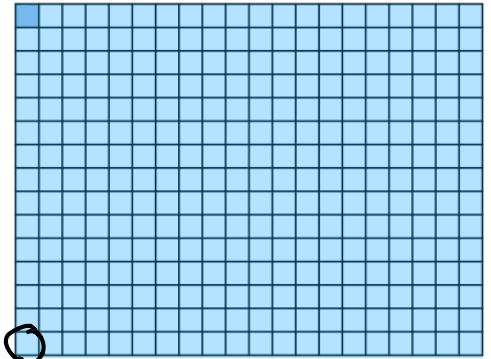
RØD



GRØNN



BLÅ



$[0, 255] \rightarrow$ skaler: deler p 255



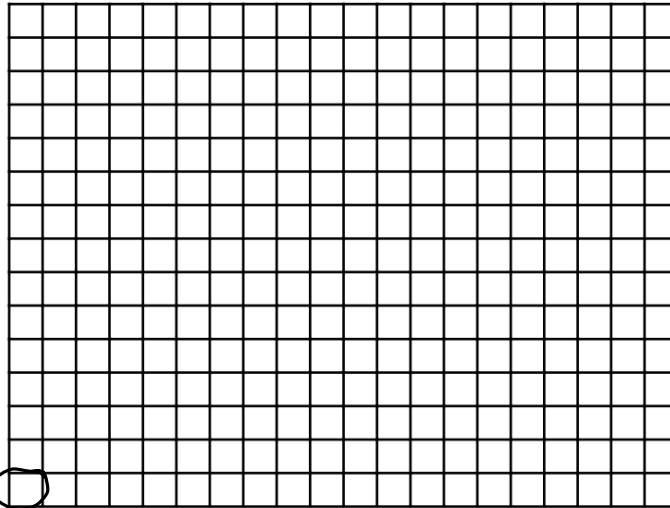
$[0, 1]$
i prosjektet

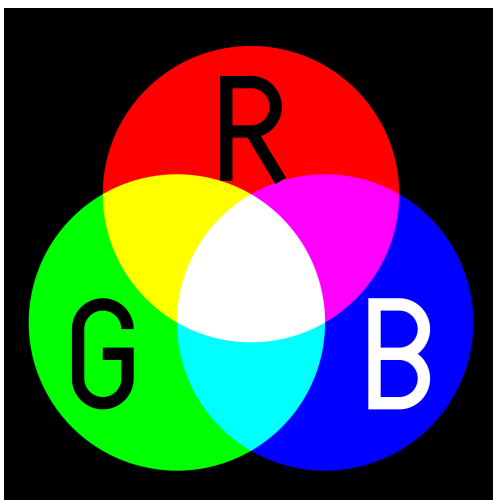
0-255

$$256 = 2^8$$

8 bit pr farge

additiv
fargeblanding





Hint: $(255, 255, 255) = (1, 1, 1)$

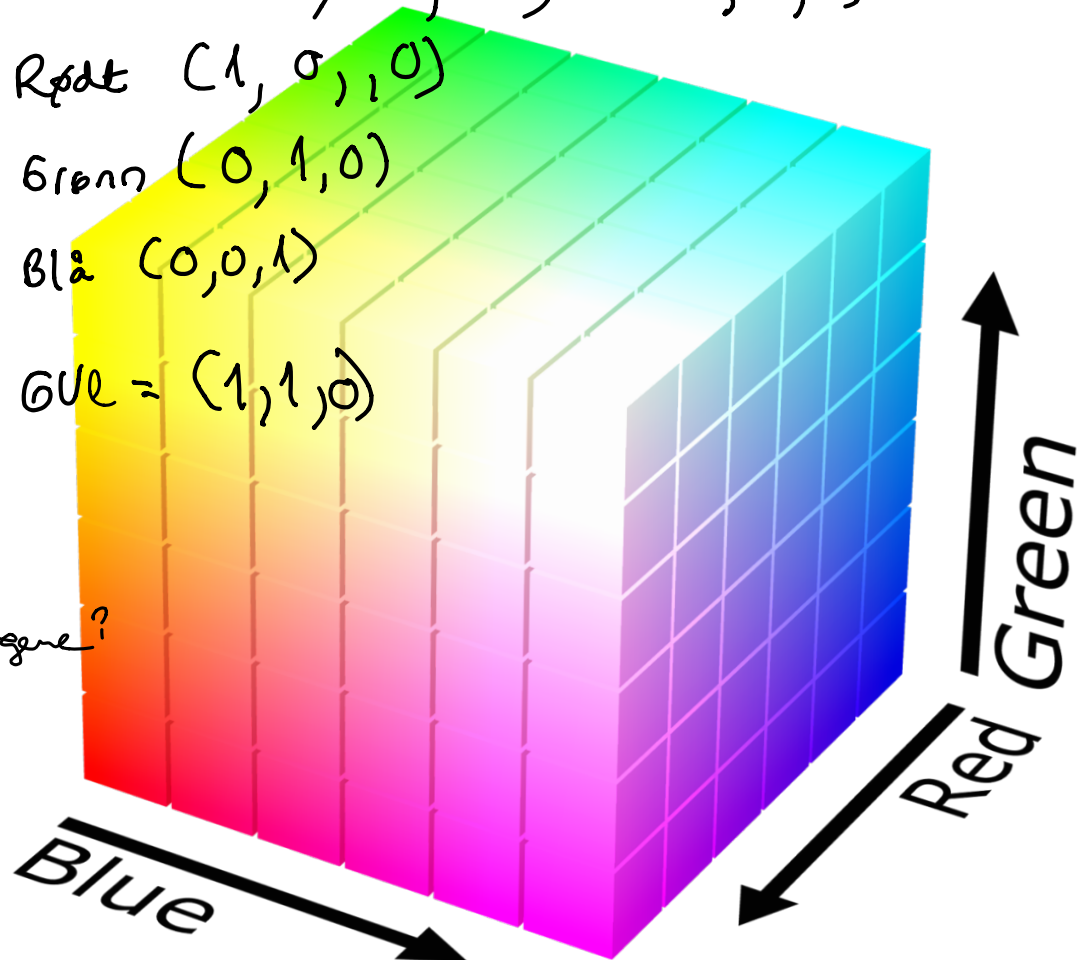
Red $(1, 0, 0)$

Green $(0, 1, 0)$

Blue $(0, 0, 1)$

GRE = $(1, 1, 0)$

↑
Hvordan
kodes farverne?



By SharkD at English
WikipediaLater
versions were
uploaded by
Jacobolus at
en.wikipedia. -
Transferred from
en.wikipedia to
Commons., Public
Domain, [https://
commons.wikimedia.or
g/w/index.php?
curid=2529435](https://commons.wikimedia.org/w/index.php?curid=2529435)

Hva er klyngeanalyse?

$X : n \times p$

~~Y~~ ← vi har ingen respons!

Ikke-veiledet = unsupervised

klasifikasjon og regresjon har vi en respons

= veiledet læring

- 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 detaljer, spare lagringsplass
- medisin : finne subgrupper av sykdom $\Rightarrow$ relevant for behandling?

Kan vi stole på resultatene? Hvor robuste er de?

fremdeles et interessant forskningsområde!

Bilde etter 2-gjennomsnitt klyngeanalyse



Resultat
← fra klyngeanalyse

Published: 17 August 2000

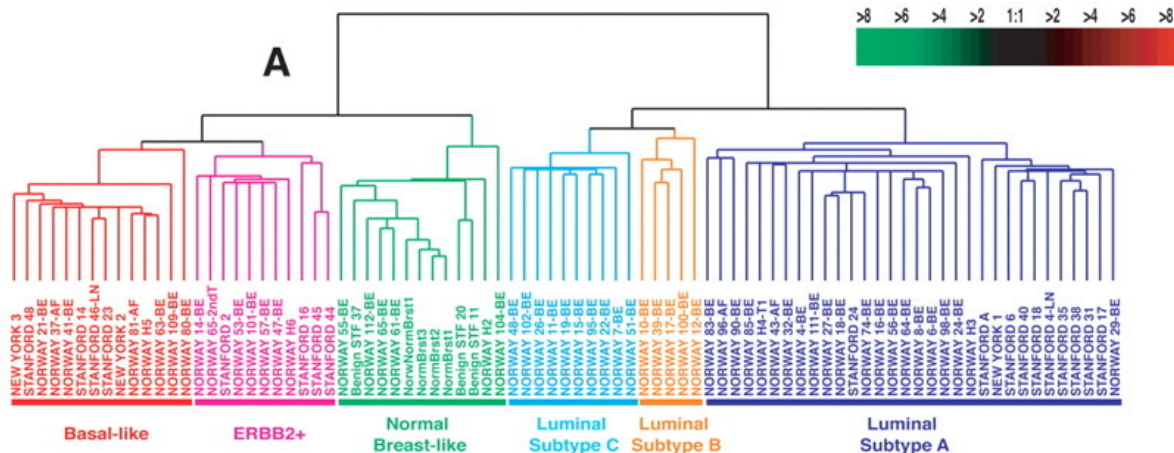
Molecular portraits of human breast tumours

Charles M. Perou, Therese Sørlie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Christian A. Rees, Jonathan R. Pollack, Douglas T. Ross, Hilde Johnsen, Lars A. Akslen, Øystein Fluge, Alexander Pergamenschikov, Cheryl Williams, Shirley X. Zhu, Per E. Lønning, Anne-Lise Børresen-Dale, Patrick O. Brown & David Botstein 

✓ collect
for Hysgene analysis

Nature **406**, 747–752(2000) | [Cite this article](#)

19k Accesses | **9375** Citations | **78** Altmetric | [Metrics](#)



Læringsmål: klyngeanalyse

- 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
- forstå hvordan klyngeanalyse utføres i Python
- **kunne besvare oppgave 3 av prosjektoppgaven på en god måte!**

Læringsressurser klyngeanalyse

Kompendium: Klyngeanalyse (pdf og html)

Videoer:

1. Klyngeanalyse (8:43 min)
2. Hierarkisk klyngeanalyse (11:26 min)
3. K-gjennomsnitt-klyngeanalyse (8:38 min)

<https://wiki.math.ntnu.no/istx1003/2020h/start>

Zoom: video og slides med notater (lenke Bb uke 46)

Pensum er definert som

**"svarene på det du blir spurt om på
prosjektoppgaven"**

og de kan du finne ved å bruke læringsressursene.

Oppgave 3: Klyngeanalyse. 6 spørsmålspunkt.

forstå hvordan
et RGB-bilde
kan bli X

tolke
kryssplott fra
RGB-bilde

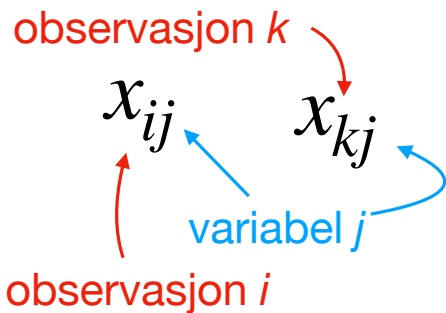
K-gjennomsnitt
klyngeanalyse
avstandsmål
sentroider
klyngemedlemskap

hierarkisk klyngeanalyse
avstandsmål
(avstand mellom klynger)
(dendrogram)

tolke resultater fra
K-gjennomsnitt
klyngeanalyse

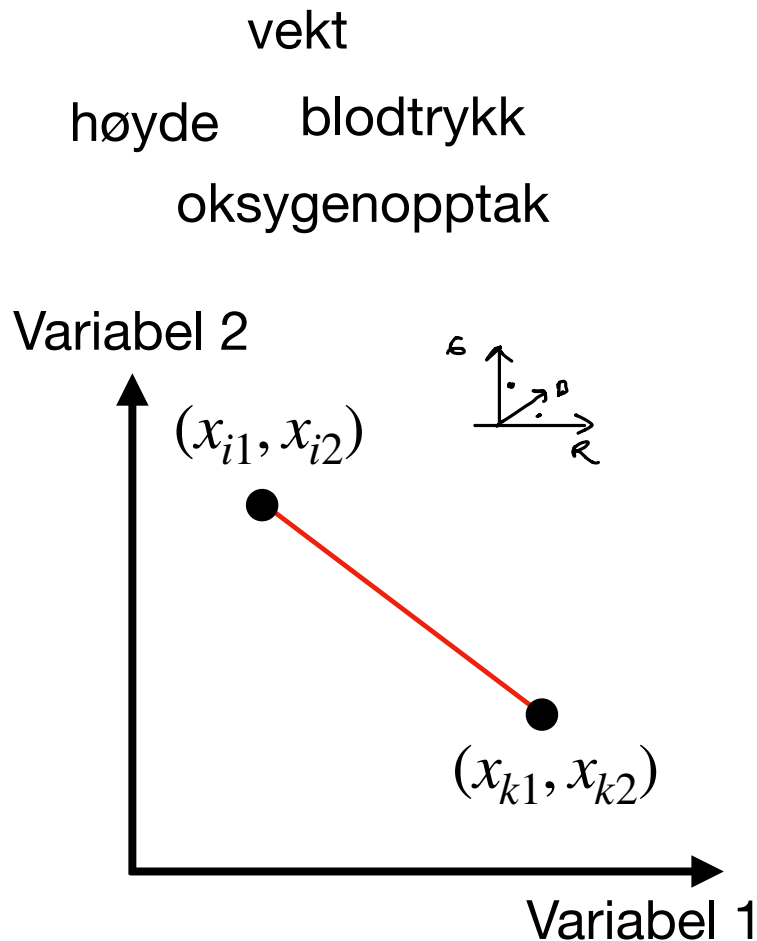
Pensum: hva er spurt om i Oppgave 3

Avstandsmål

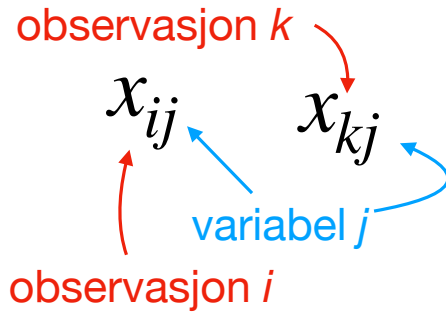


Euklidsk avstand

$$D_E(i, k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2}$$



Avstandsmål

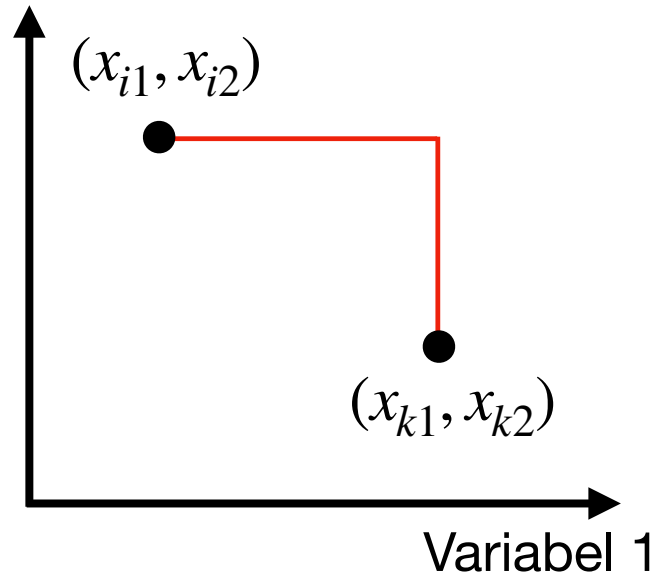


City Block-avstand

aka Manhattan

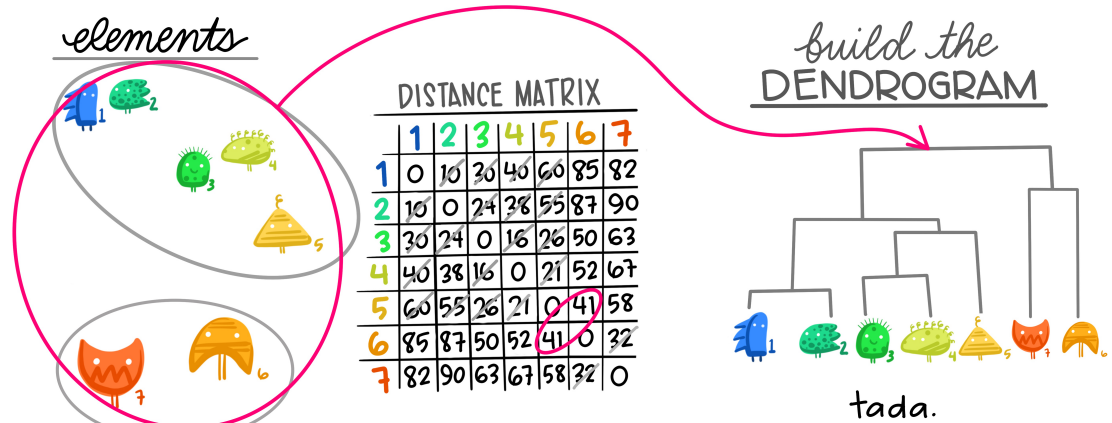
$$D_M(i, k) = \sum_{j=1}^p |x_{ij} - x_{kj}|$$

Variabel 2



Videre: metoder for klyngeanalyse

Hierarkisk
klyngeanalyse



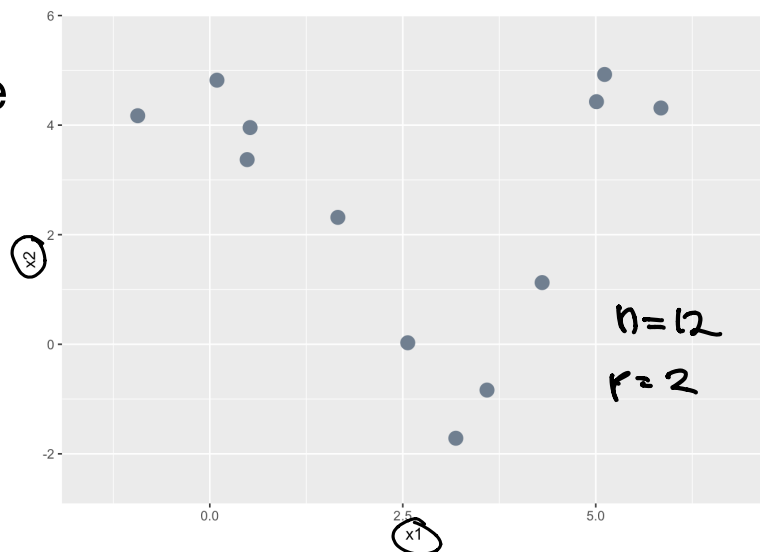
K-gjennomsnitt
klyngeanalyse

K-gjennomsnitt klyngeanalyse

- Finn ukjente klynger i dataene: K klynger

$n \hat{=}$ velger

- *alle* observasjoner skal være medlem i en klynge
- hver observasjon kan bare være medlem i *en* klynge
- variasjonen innen hver klynge skal være så *liten* som mulig



urealistisk eksempel for å vise algoritmen

Kvadrert euklidsk avstand brukes som avstandsmål.

Variasjon innen en klynge k

$$C_k = \{1, 2, 3\}$$

Hvilke observasjoner er med i klynge k

observasjon i'

x_{ij}

$x_{i'j}$

variabel j

observasjon i

$$|C_k| = 3$$

antall observasjoner i klynge k

asterisker teller

$$10^2$$

$$10^3$$

$$2 \cdot 10^3$$

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

Antall mulige løsninger

$$= K \cdot K \cdot \dots \cdot K = K^n$$

$$n = 100, p = 3$$

$$3^{100} \sim 5 \cdot 10^{45}$$

Optimeringsproblem

Minimer over alle

$$C_1, \dots, C_K$$

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

kan ikke
sjekke alle
løsningene
og velg
den
beste
→ forferkelig

nyttig sammenheng som er grunnlaget
for k-gjennomsnitt algoritmen

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

$$\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$$

Klyngesentroide i klynge k: $(\bar{x}_{k1}, \bar{x}_{k2}, \dots, \bar{x}_{kp})$

↑
Viktig kjernebegrep!

K -gjennomsnitt-algoritmen

Klyngesentroide: $(\bar{x}_{k1}, \bar{x}_{k2}, \dots, \bar{x}_{kp})$

- Start med å velge K
- Tilordne hver observasjon til en klynge
 - Mange muligheter, en er å tilfeldig velge ut K observasjoner og sette disse som klyngesentroider
 - og deretter tilordne der resterende observasjoner til klyngen med nærmeste klyngesentroide

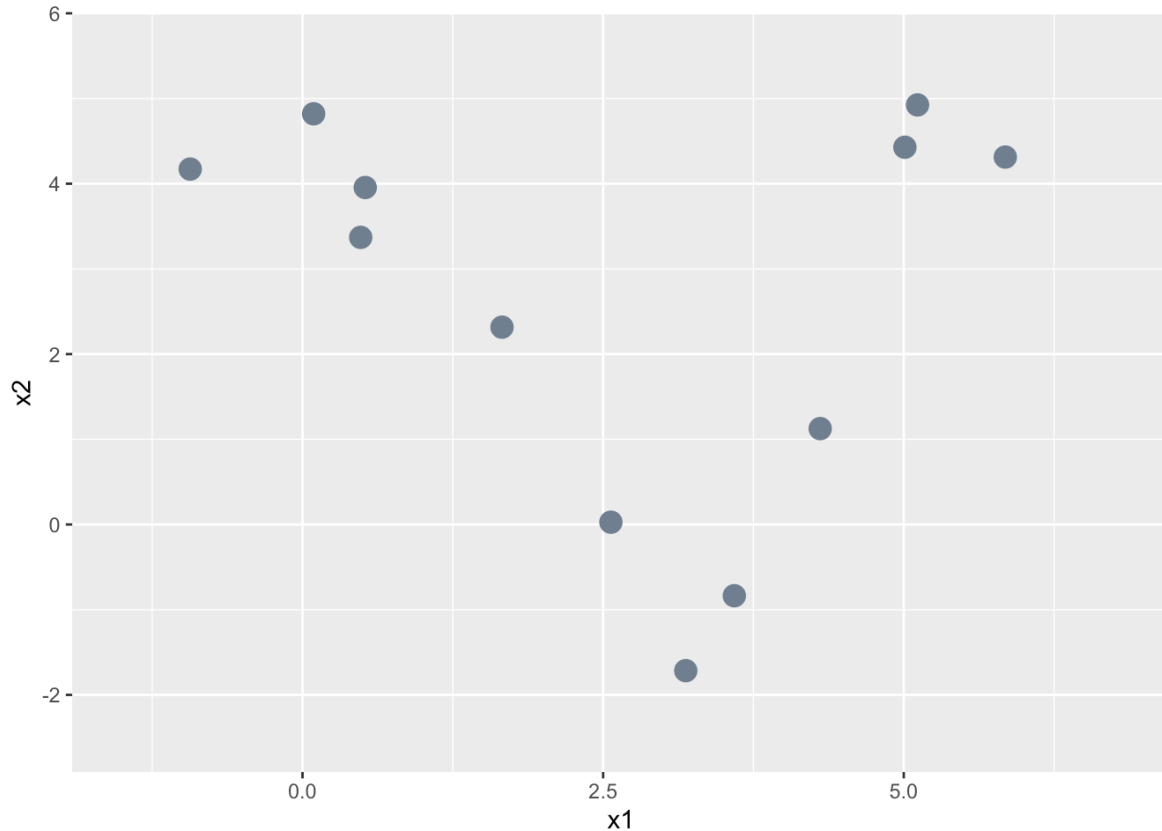
Repeter til ingen observasjoner endrer klyngemedlemskap:

1. For hver klynge regn ut klyngesentroiden
2. Tilordne hver observasjon til klyngen til nærmeste klyngesentroide

K -gjennomsnitt-algoritmen i praksis

$n=12$ og $p=2$

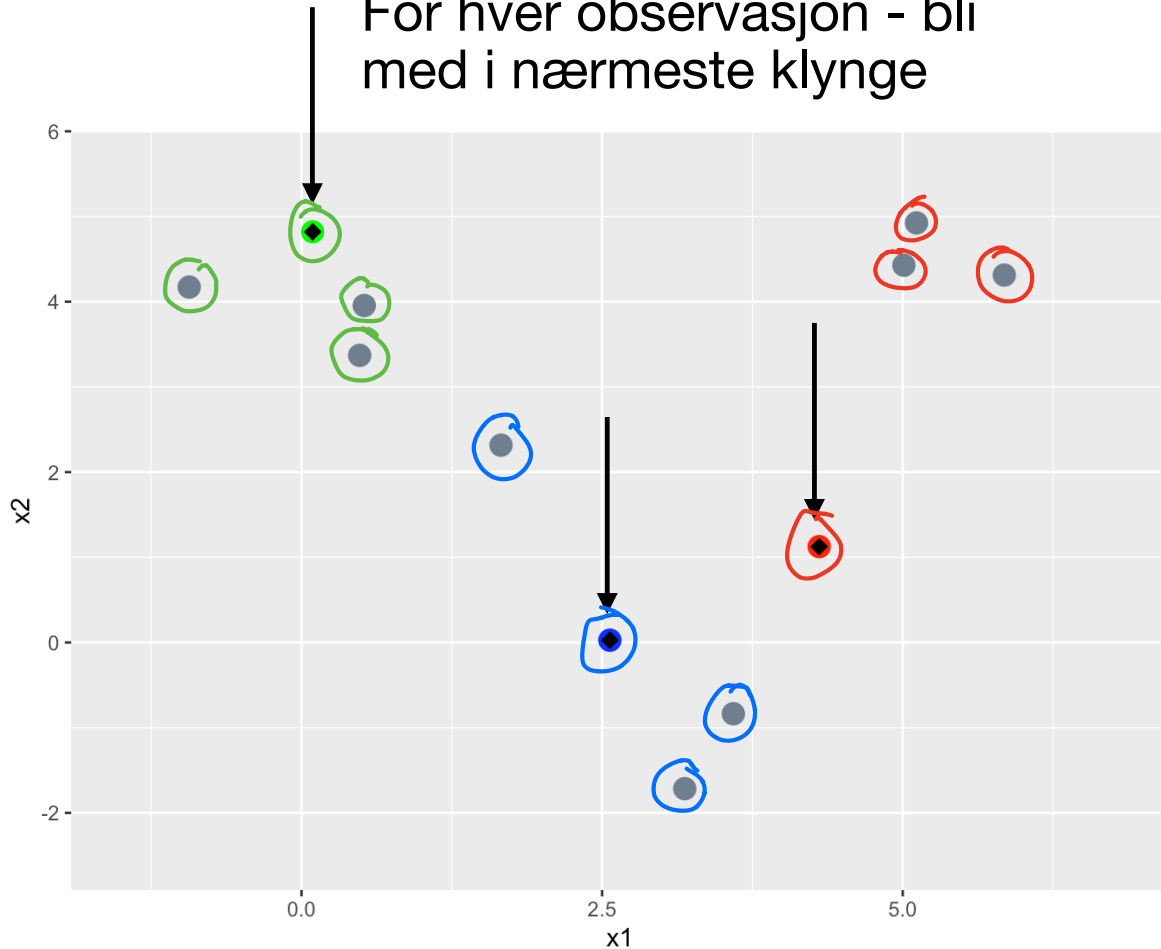
Finn $K=3$ klynger



Initialisering

Velger tilfeldig ut $K=3$ observasjoner som klyngesentroider

For hver observasjon - bli med i nærmeste klynge



K -gjennomsnitt-algoritmen

Klyngesentroide: $(\bar{x}_{k1}, \bar{x}_{k2}, \dots, \bar{x}_{kp})$

- Start med å velge K
- Tilordne hver observasjon til en klynge
 - Mange muligheter, en er å tilfeldig velge ut K observasjoner og sette disse som klyngesentroider
 - og deretter tilordne der resterende observasjoner til klyngen med nærmeste klyngesentroide

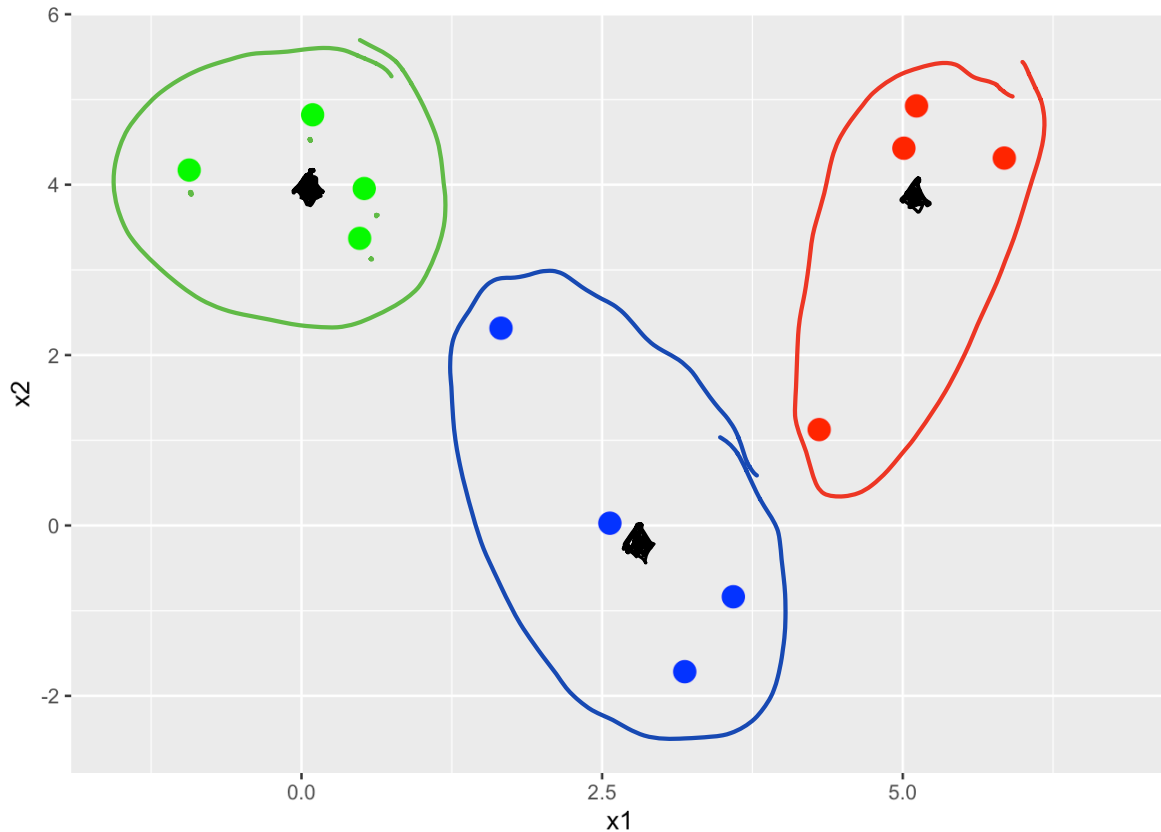
Repeter til ingen observasjoner endrer klyngemedlemskap:

1. For hver klynge regn ut klyngesentroiden
2. Tilordne hver observasjon til klyngen til nærmeste klyngesentroide

Repeter til ingen observasjoner endrer klyngemedlemskap

Regn ut nye klyngesentroider

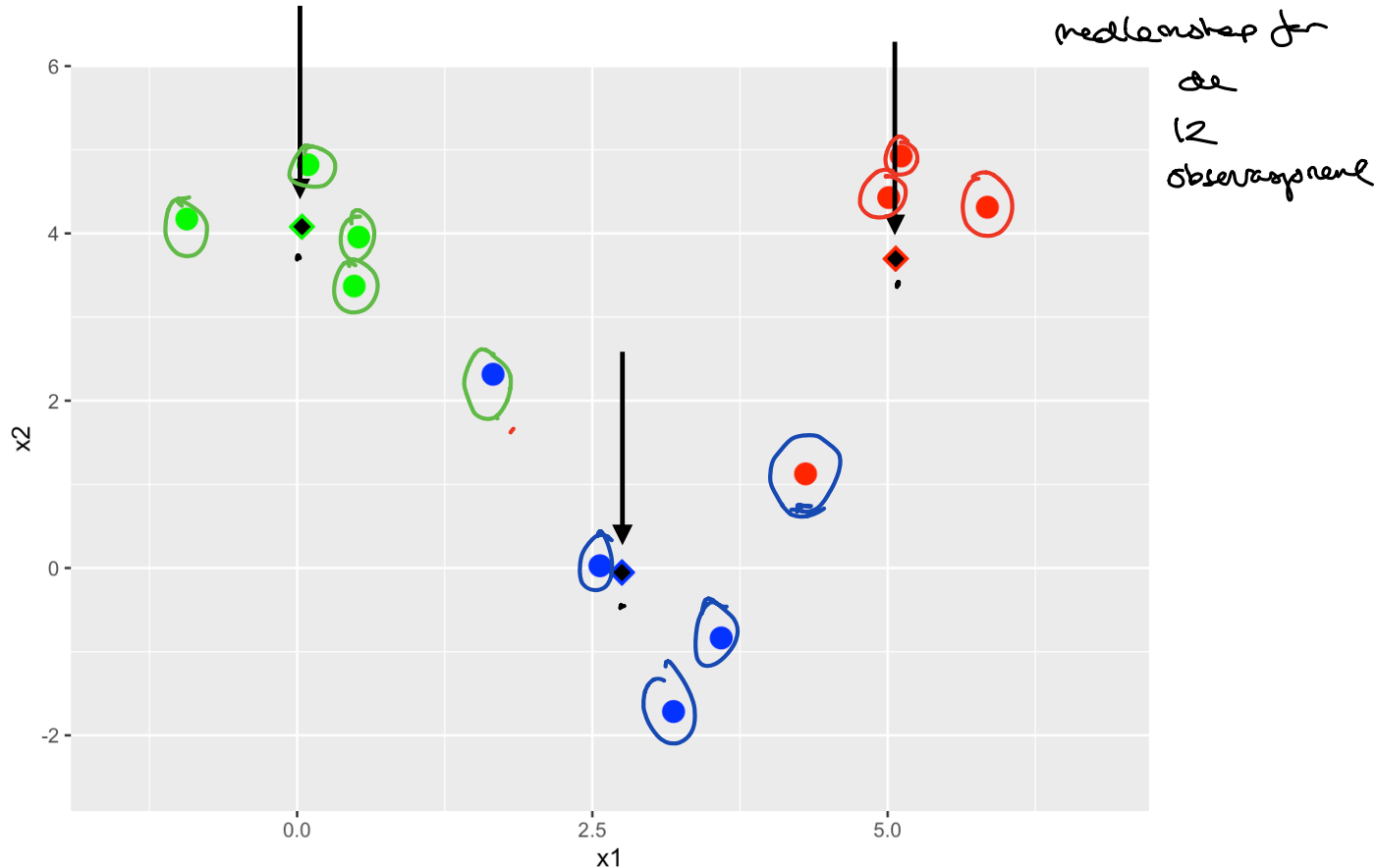
$$\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$$



◆ nye
sentroider
for
hver
klynge

Nye klyngesentroider

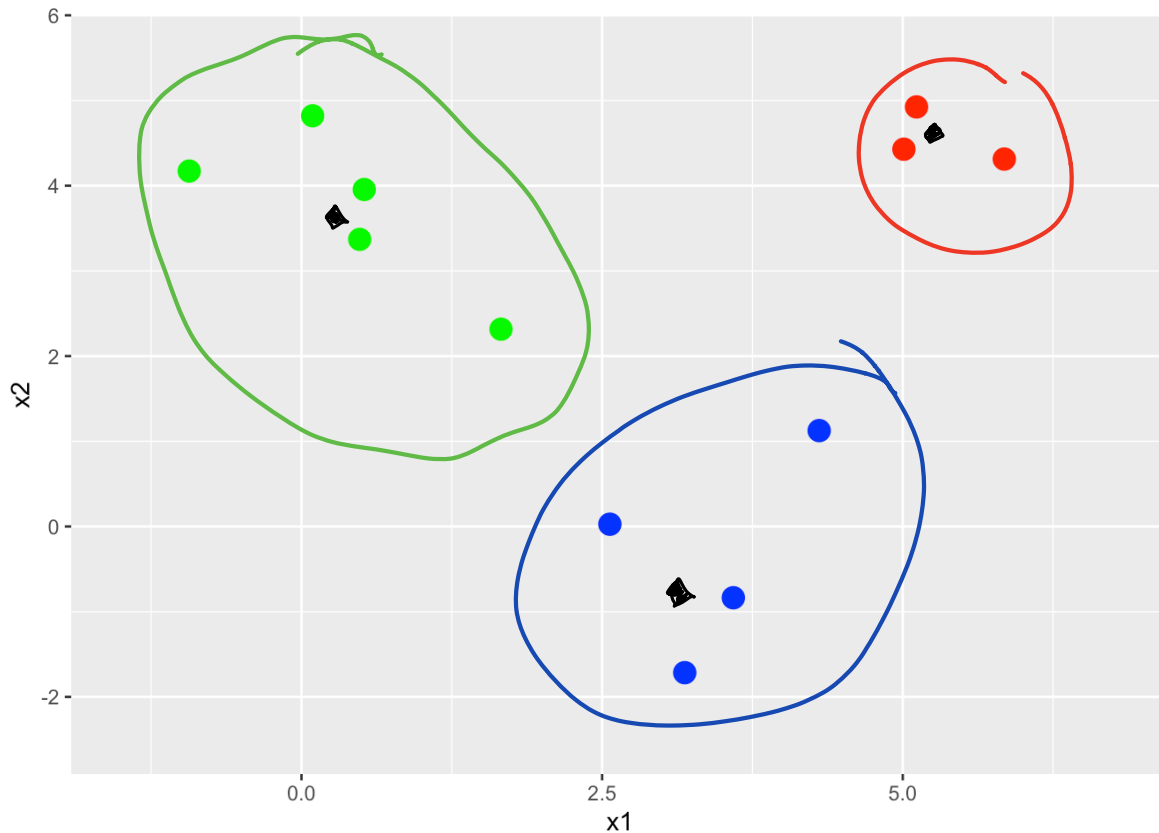
For hver observasjon - bli med i nærmeste klynge



Endret noen observasjoner klyngemedlemskap? ~~JA~~ → fortsett

Repeter til ingen observasjoner endrer klyngemedlemskap

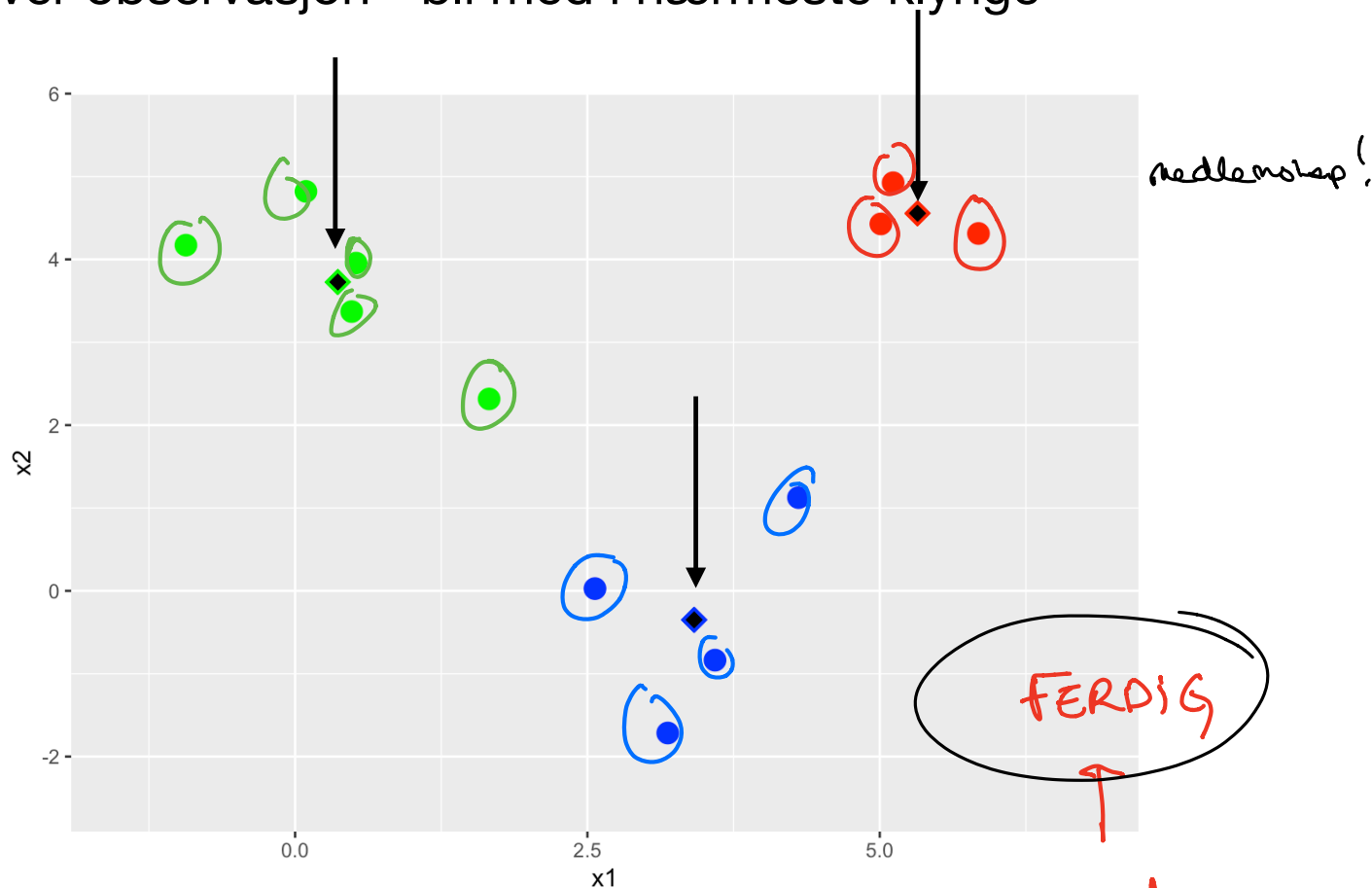
Regn ut nye klyngesentroider



◆ nye
sentroider

Nye klyngesentroider

For hver observasjon - bli med i nærmeste klynge



Endret noen observasjoner klyngemedlemskap?

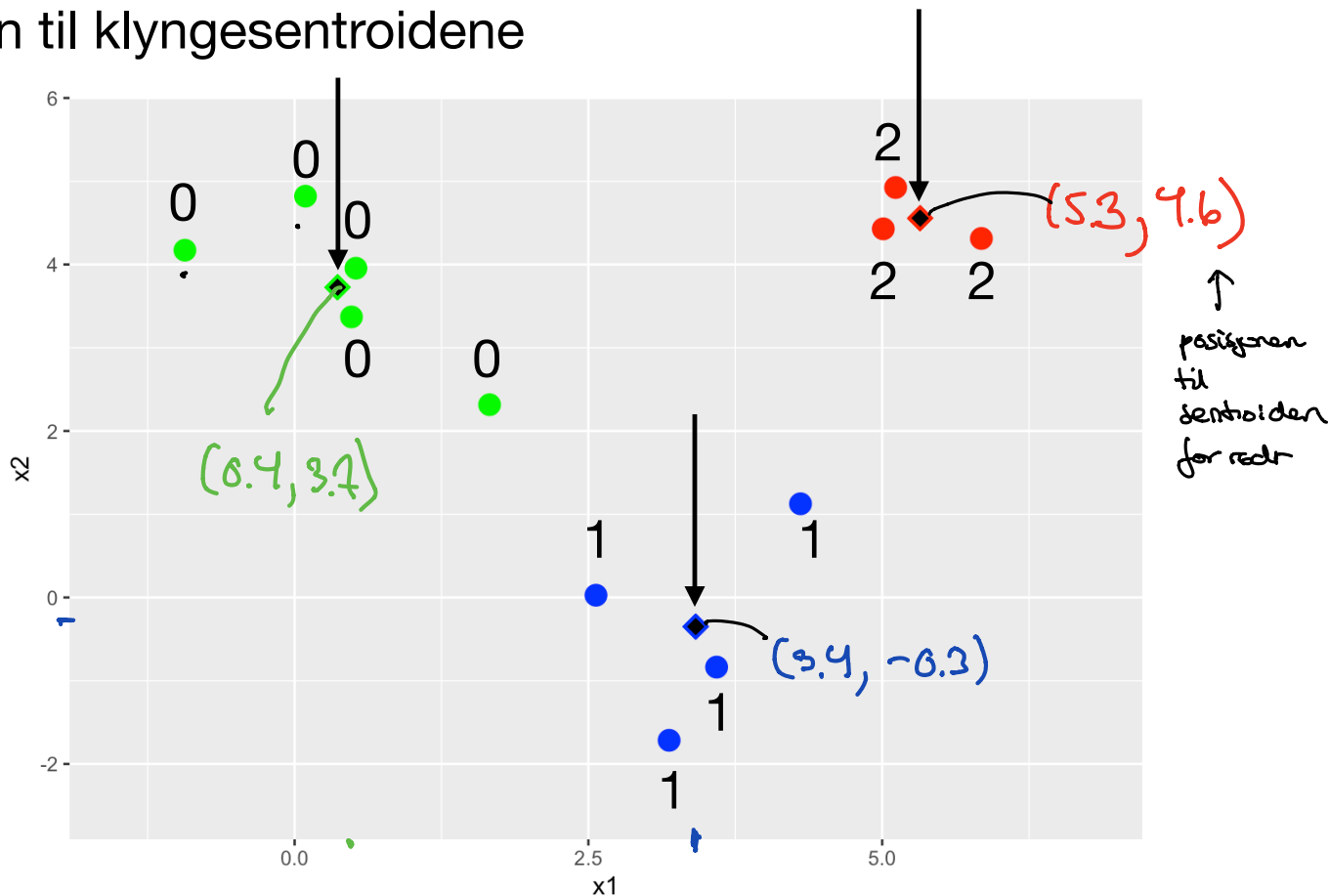
NEI

Resultat

// også kalt "klyngemerkebeffer"

Klyngemedlemskap for hver observasjon

Posisjon til klyngecentroidene



K-gjennomsnitt-algoritmen i Python

$\sum_{n \times p}$

→ df: datasett med antall observasjoner
ganger antall variabler

```
from sklearn.cluster import KMeans
```

Modellvalg →

```
modell = KMeans(n_clusters=3, seed for a reproducible  
random_state=0,
```

```
n_init=10, ← antall start-  
punkter
```

```
init='k-means++' ← smart sørgen av  
initialisering)
```

Tilpass modellen
til data →

```
tilpass=modell.fit(df)
```

Klyngemedlemskap →

```
data_klynge=tilpass.labels_ (... 0, 1, 1, 2, 0, ...)
```

Klyngesentroider →

```
sentroider=tilpass.cluster_centers_  
(0.4, 3.7), (3.4, -0.3), (5.3, 4.6)
```

Sentroiden til
hver observasjon →

```
data_sentroider=sentroider[data_klynge]  
(... (0.4, 3.7), (3.4, -0.3), (3.4, -0.3), (5.3, 4.6), (0.4, 3.7), ...)
```

A photograph of the Leifeng Pagoda in Hangzhou, China. The pagoda is a multi-tiered traditional Chinese structure with red walls, green roofs, and ornate carvings. It is situated on a hill overlooking a large body of water, with trees in the foreground and a distant city skyline visible on the horizon.

[illegible]

← $\Sigma =$
data_finger

Prosjektoppgaven 3c)

`data_farger` er en n ganger p matrise med "bildet"

```
# Steg 1: Spesifiser antall klynger
```

```
antall_klynger = 2
```

```
# Steg 2: Initialiser k-means algoritmen
```

```
kmeans = KMeans(n_clusters = antall_klynger)
```

legg til n -init
randomskole

```
# Steg 3: Tilpass modellen
```

```
kmeans.fit(data_farger)
```

← her utføres algoritmen og de leges sentroide klynge medelverdi

```
data_klynge=kmeans.labels_  
kmeans.predict(data_farger)
```

} gir sannersvar ($\dots 0, 1, 0, 0, \dots$)
medlemskap/medlemskap

```
sentroider=kmeans.cluster_centers_
```

← (R, G, B)
 $((R_0, G_0, B_0), (R_1, G_1, B_1))$
 R, G, B i klynge 0

```
data_sentroide=sentroider(data_klynge)
```

$(\dots (R_0, G_0, B_0), (R_1, G_1, B_1), (R_0, G_0, B_0) \dots)$

Prosjektoppgaven 3d)

Hva er hovedforskjellene mellom K -gjennomsnitt-klyngeanalyse og hierarkisk klyngeanalyse?

Vi ber deg ikke om å finne klynger i bildet ved hjelp av hierarkisk klyngeanalyse. Hva kan være grunnen til at vi ikke gjør det?

forstå hvordan
et RGB-bilde
kan bli X

tolke
kryssplott fra
RGB-bilde

K-gjennomsnitt
klyngeanalyse
avstandsmål
sentroider
klyngemedlemskap

hierarkisk klyngeanalyse
avstandsmål
(avstand mellom klynger)
(dendrogram)

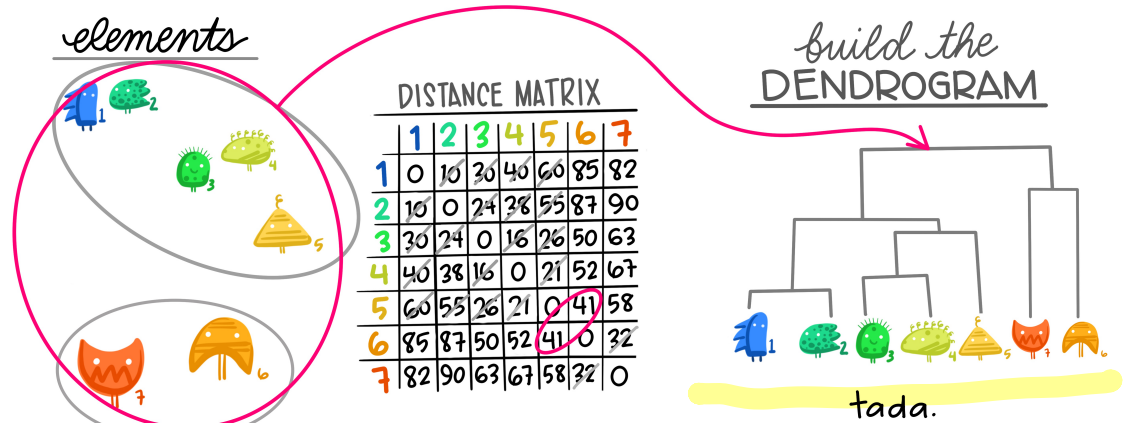
W²_A!

tolke resultater fra
K-gjennomsnitt
klyngeanalyse

Pensum: hva er spurt om i Oppgave 3

Videre: metoder for klyngeanalyse

Hierarkisk
klyngeanalyse

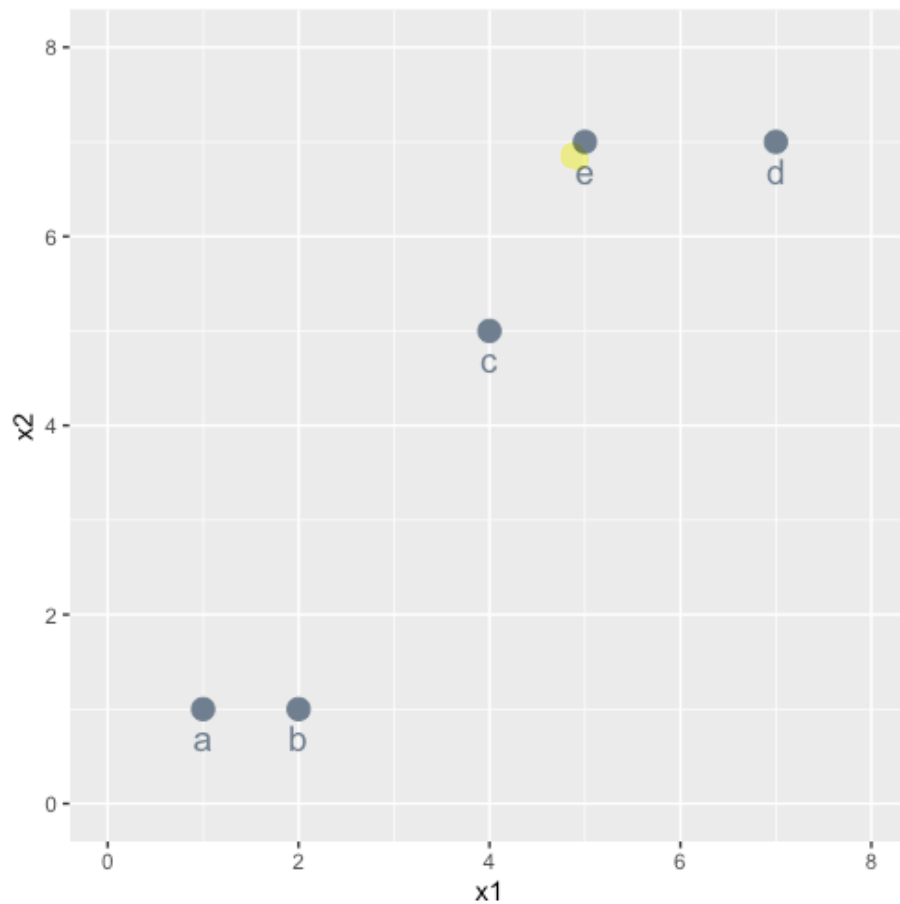


K-gjennomsnitt
klyngeanalyse

Eksempel

$$n=5, p=2$$

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

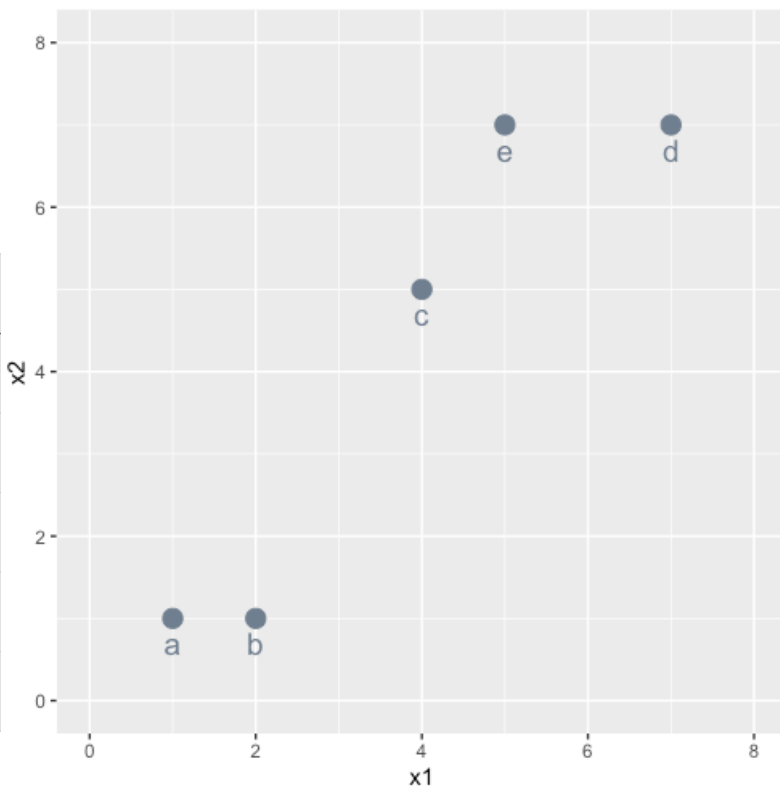


Avstandsmatrise

n observasjoner

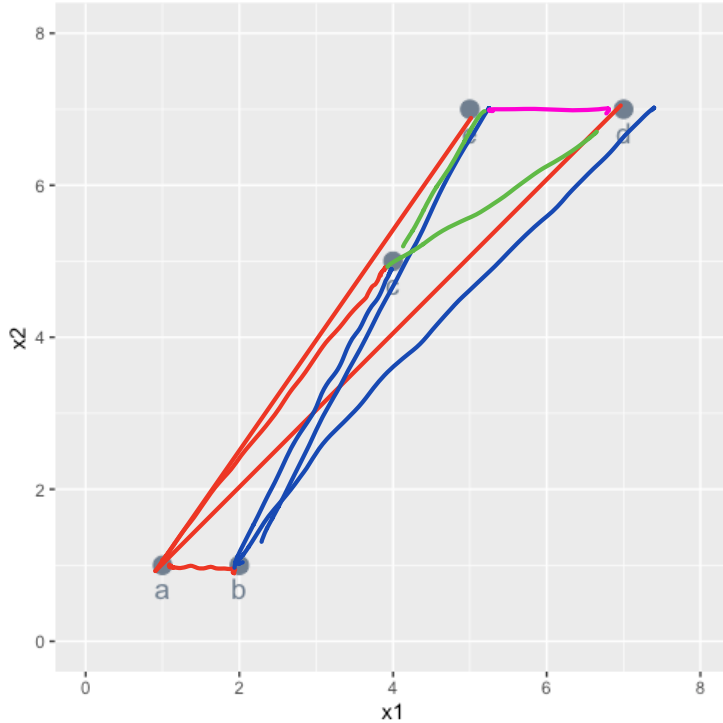
p variabler

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



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

Avstandsmatrise



regner ut 10
avstander

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

Euklidisk
avstand

symmetrisk
matrise

	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

Hvordan beregne avstand mellom klynger?

Singel kobling:

minimal avstand *b-c*

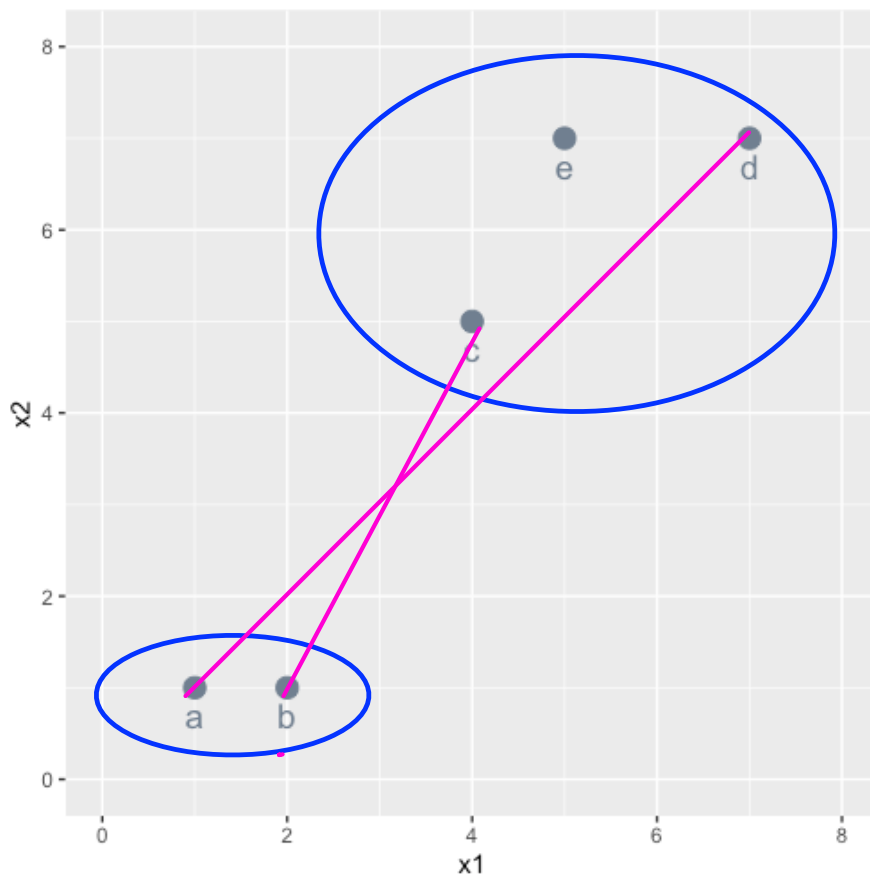
Komplett kobling:

maksimal avstand *a-d*

Gjennomsnittskobling:

gjennomsnittlig avstand

as de 6 mulige
avstandene
a-d, a-e, a-c
b-d, b-e, b-c



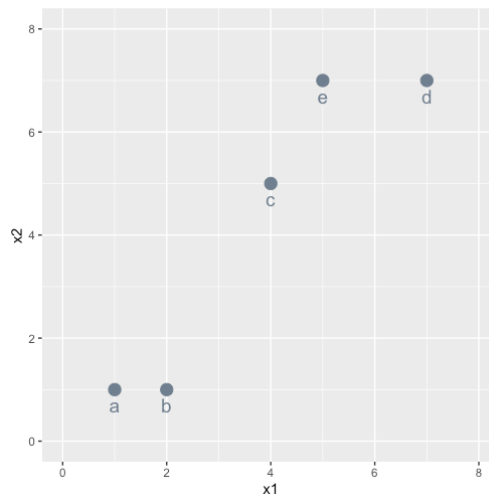
Algoritme for hierarkisk klyngeanalyse

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.

	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



Algoritme for hierarkisk klyngeanalyse

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

1. Slå sammen

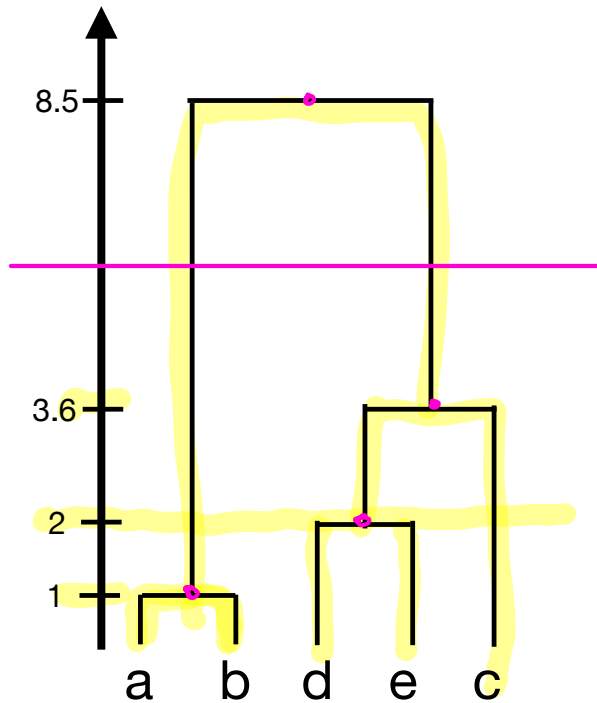
Finn de to klyngene som er nærmest hverandre og slå dem sammen til en klynge.

2. Beregn avstander

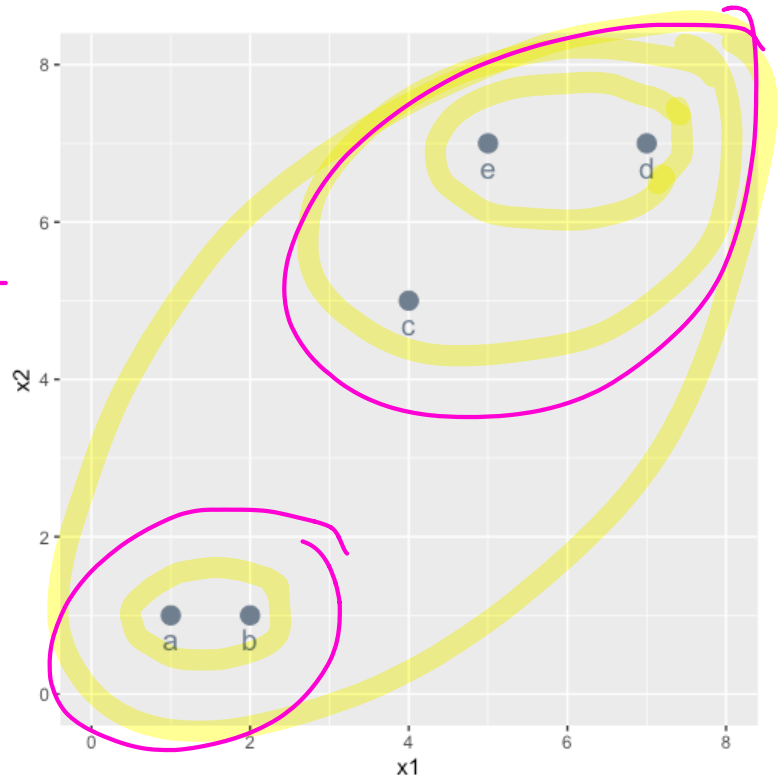
Beregn nye parvise avstander mellom alle klynger ved bruk av valgt avstandsmål og koblingstype.

Repeterer til alle observasjonene er i samme klynge.

Resultat fra hierarkisk klyngeanalyse



Dendrogram

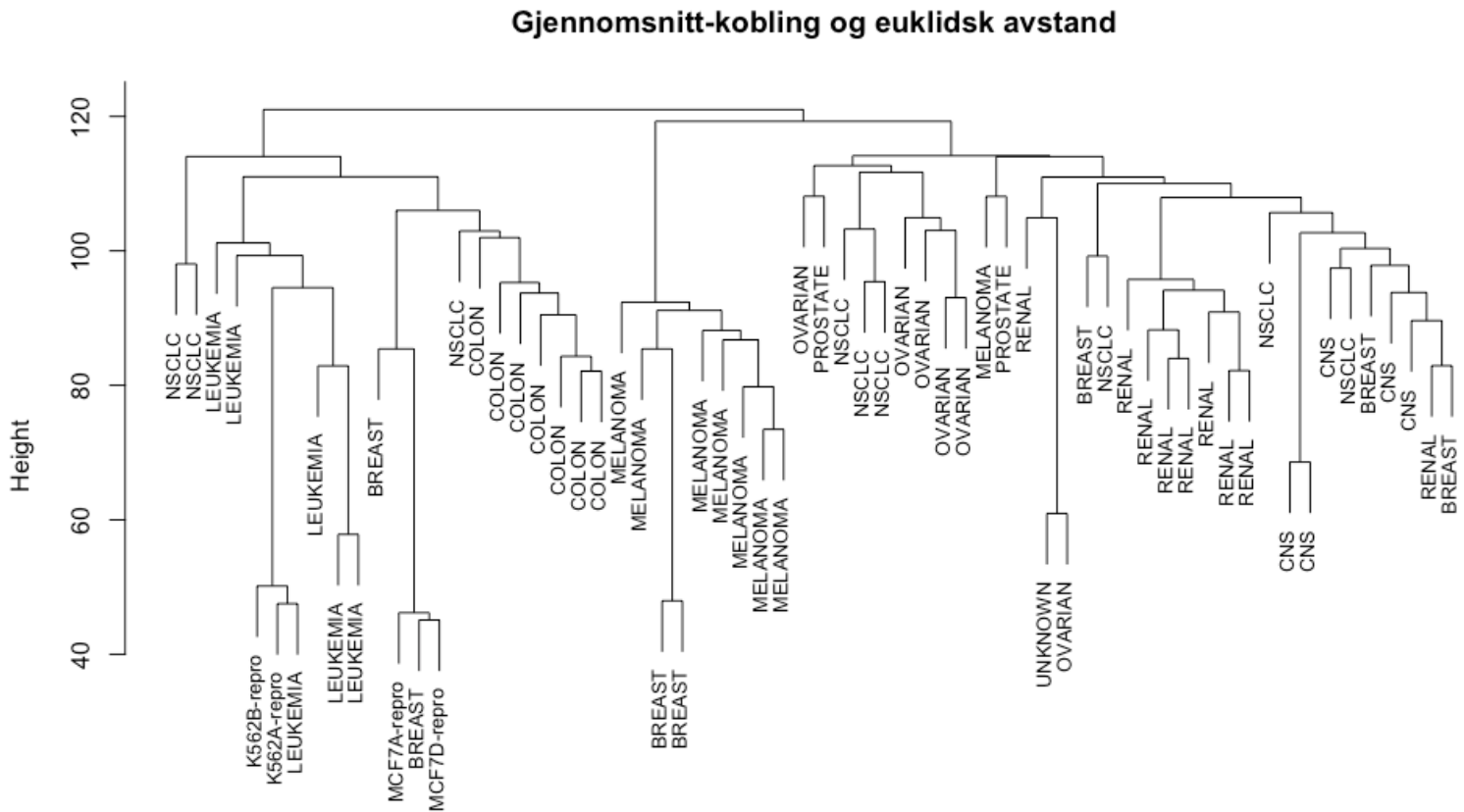


Sokkeerzoner → 4 "giftetial"

Trenger ikke bestemme k p3 forhand!

Eksempel: genaktivitet

- n=64 celleprøver fra kreftsvulster til ulike pasienter
- genaktivitet for p=6830 gener



Prosjektoppgaven 3d)

Hva er hovedforskjellene mellom K -gjennomsnitt-klyngeanalyse og hierarkisk klyngeanalyse?

Vi ber deg ikke om å finne klynger i bildet ved hjelp av hierarkisk klyngeanalyse. Hva kan være grunnen til at vi ikke gjør det?

Er dette nok til at du kan svarer på spørsmålene?

Nei? Da må du se videoen: “Hierarkisk klyngeanalyse” og lese i kompendiet!

gjensker

forstå hvordan
et RGB-bilde
kan bli X

tolke
kryssplott fra
RGB-bilde

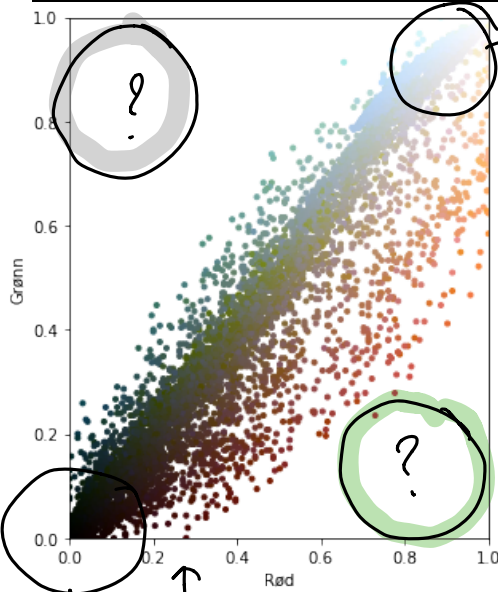
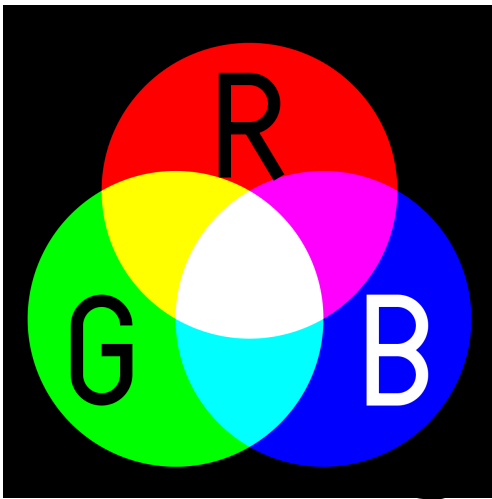
K-gjennomsnitt
klyngeanalyse
avstandsmål
sentroider
klyngemedlemskap

hierarkisk klyngeanalyse
avstandsmål
(avstand mellom klynger)
(dendrogram)

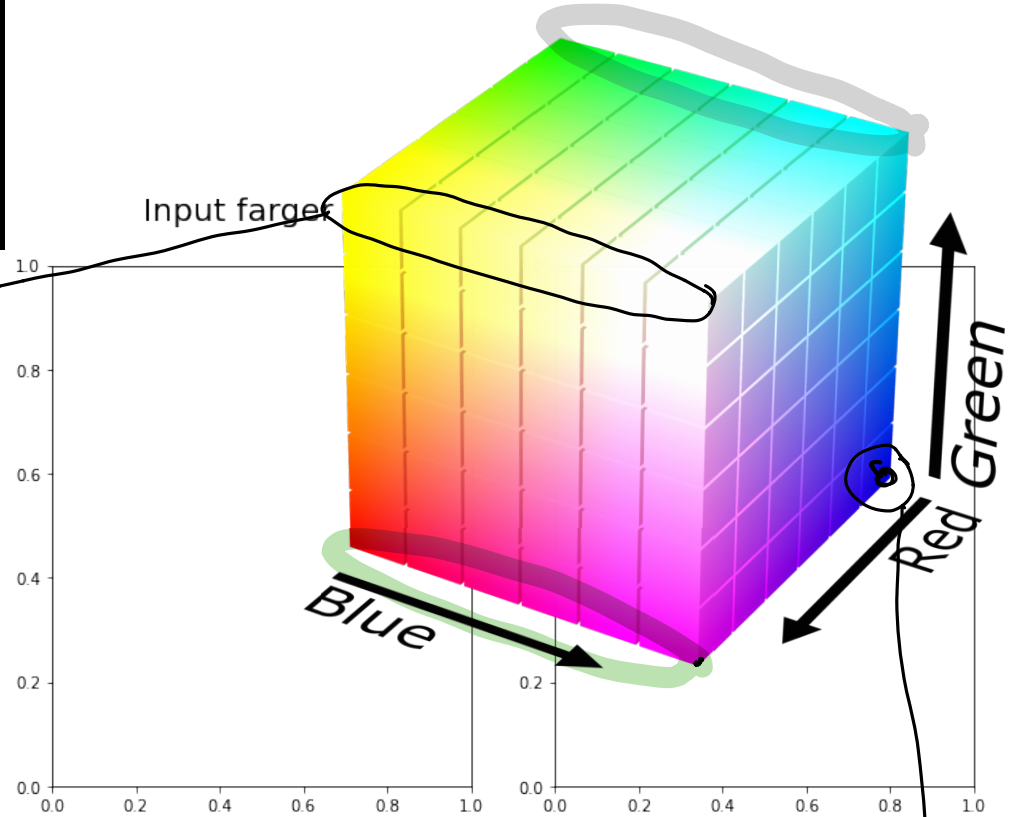
tolke resultater fra
K-gjennomsnitt
klyngeanalyse

Pensum: hva er spurt om i Oppgave 3

Prosjektoppgaven 3b og 3d



↑
en piksel = et piksel



Plan: Klyngeanalyse

1. Hva er klyngeanalyse?
2. Læringsmål, pensum og læringsressurser
3. Avstandsmål
4. K-gjennomsnitt klyngeanalyse
5. Bruk av klyngeanalyse på et bilde (prosjektet)
6. Hierarkisk klyngeanalyse
7. Har vi dekket pensum?
8. Innspurt prosjektmodulen mot 29.11 kl 23.59

(Inkludert 15 minutters pause)

Innlevering av prosjektet

- Prosjektoppgaven er laget som Jupyter-notatbok - og finnes i det formatet på hubben, og også som html for å bare lese (ikke kjøre cellene).
- KORTE svar på spørsmål skrives i separat dokument: vise malen!
- Svarene på spørsmålene skriver dere inn i malen og lagrer i pdf-format.
- Noen oppgaver krever modifikasjon av Jupyter-notatbok - den modifiserte notatboken lagrer du i ipynb-format: vise hvordan!
- Pdf og modifisert ipynb lastes opp på Bb i prosjektinnleveringen: vise hvor!
- Innleveringsfrist 29.11.2020 kl 23.59 på Bb.
- Gis bokstavkarakter og teller 20% av sluttkarakter.

Videre denne og de neste ukene

- Ta med dere gruppa deres på digital veiledning!
- Jobb med alle oppgavene på prosjektet!
- Still korte spørsmål i Forumet
- og lange spørsmål på digital veiledning mandag 8.15-10 og tirsdag 12.15-14 + egne tider Gjøvik.
- Neste uke er siste undervisningsuke - med oppsummeringstimer på zoom fra Fellesmodulen - med Thea!

Læringsmål

Etter du har gjennomført denne modulen skal du kunne:

- forstå når du kan bruke regresjon, klassifikasjon og klynganeanalyse til å løse et ingeniørproblem
- – spesifikke mål for hver av de tre temaene –
- bli en kritisk leser av resultater fra statistikk/maskinlæring/statistisk læring/data science/kunstig intelligens når disse rapporteres i media, og forstå om resultatene er realistiske ut fra informasjonen som gis
- **kunne besvare prosjektoppgaven på en god måte!**

Lykke til!