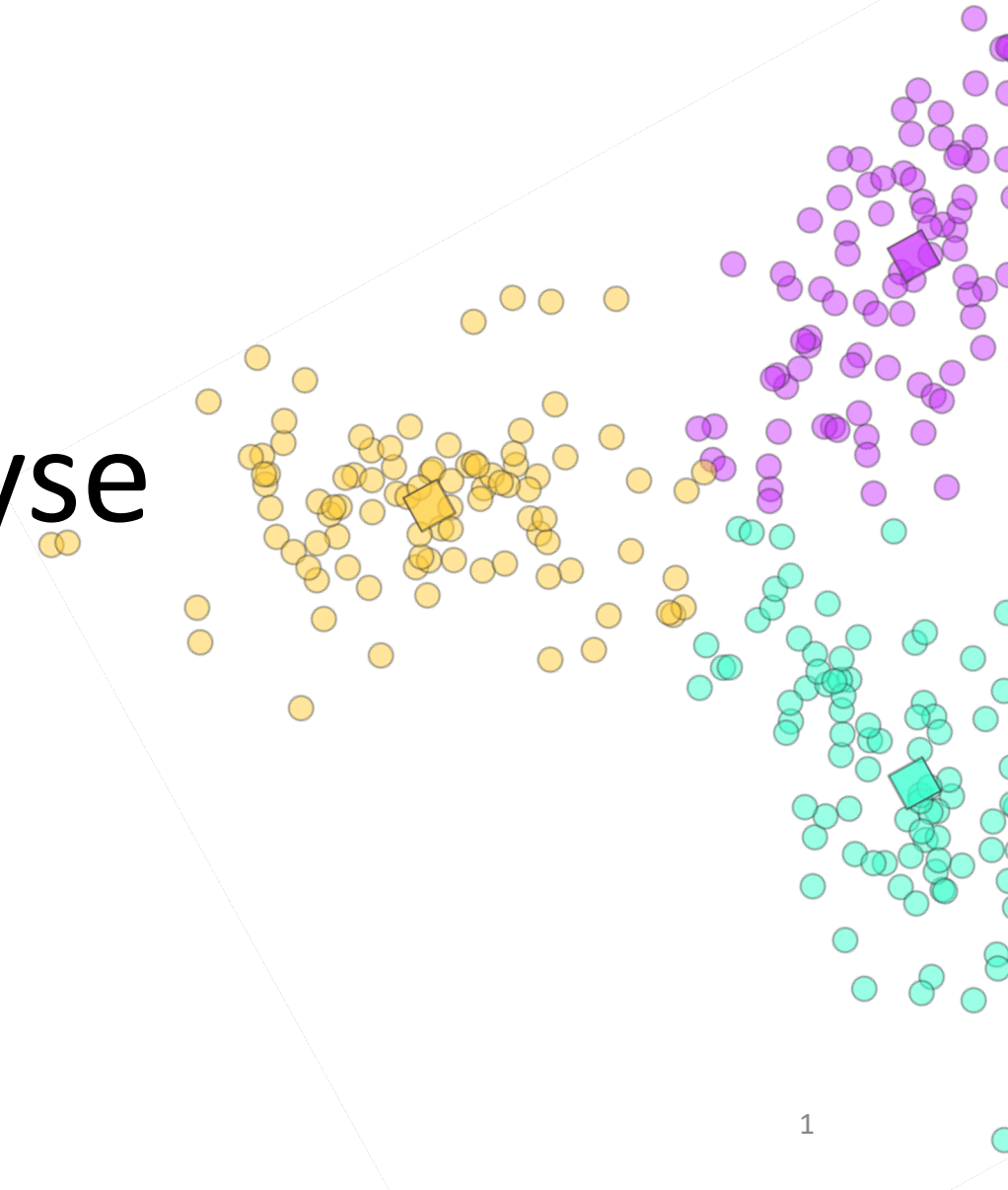
Klyngeanalyse



Klyngeanalyse

- Ikke-veiledet maskinlæring

- Deler dataen inn i klynger, slik at like objekter er i samme klynge

- Bruker Euklidsk avstand som likhetsmål

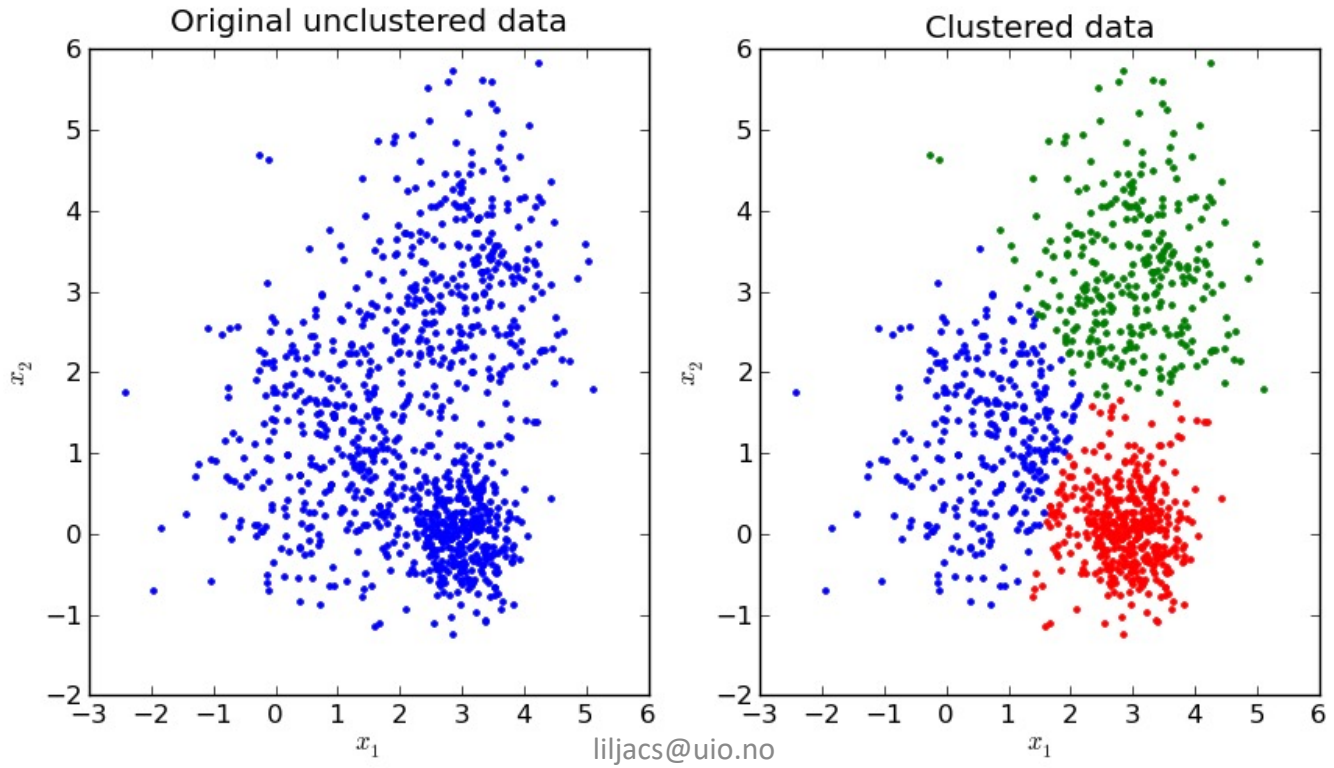
- Hierarkisk clustering: lager en trestruktur av hierarkisk nøstede klynger

- Flat clustering: Forsøker å direkte dekomponere data til et sett av klynger

k-means clustering

Ikke-veiledet versjon av Rocchio

Mål: Dele inn dataen i klynger, slik at hvert punkt hører til klyngen med nærmeste centroide



WCSS: *Within-cluster sum of squares.* Tapsfunksjon som måler hvor godt hver centroide representerer medlemmene av klyngen

$$WCSS = \sum_{c_i \in C} \sum_{x_j \in c_i} ||x_j - \mu_i||^2$$

Med andre ord: Vi vil minimere det gjennomsnittlige kvadratet av distansen mellom objekter og deres klyngecentroider

Jo høyere tall, desto mer «bommer» modellen vår.

Algoritmen

Initialiser: velg *k* tilfeldige «seeds»

Iterer:

- Tilordne hvert objekt den klyngen med nærmeste centroide

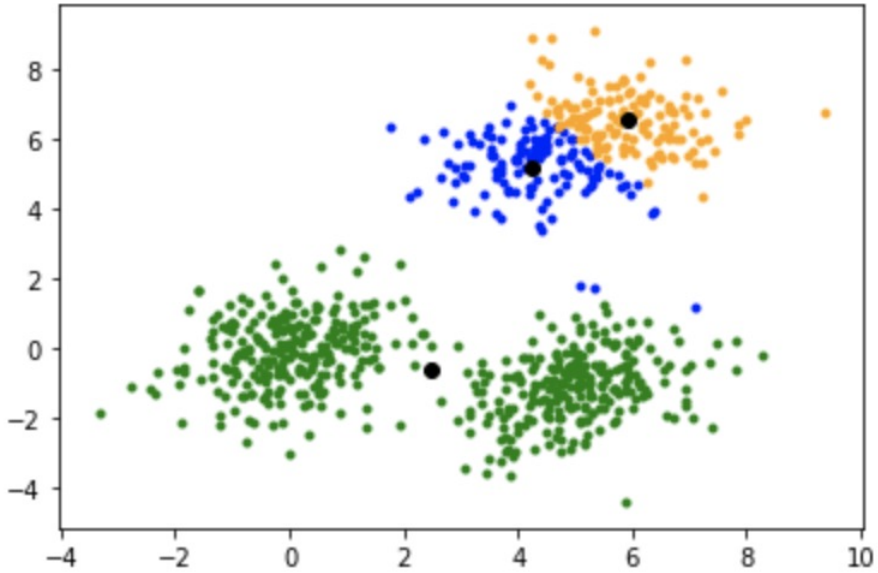
- Generer nye centroider for klyngene

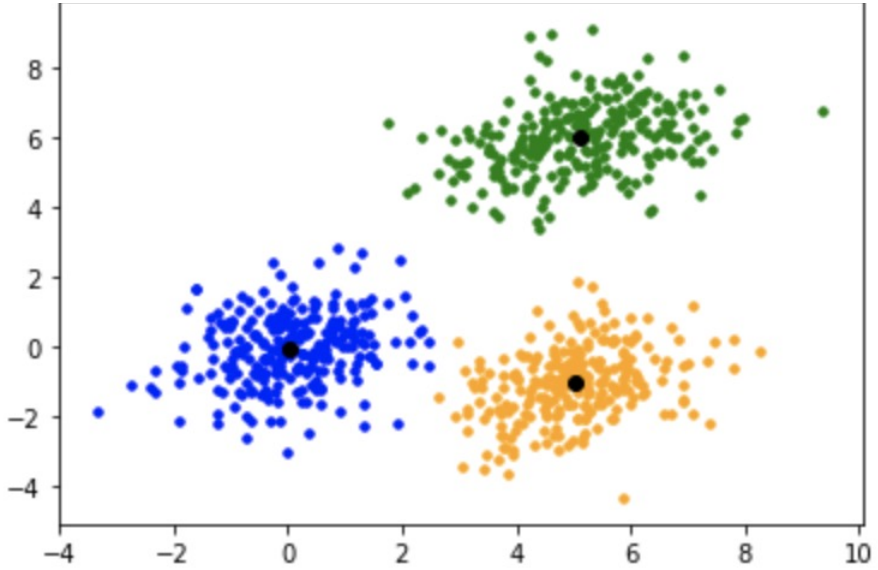
Terminer: Når vi har oppfylt stoppkriteriet vårt

Hvordan vet vi når vi skal stoppe?

Vi må et stoppkriterie:

- Et forhåndsbestemt antall iterasjoner
- Når centroidene ikke forandrer seg mer mellom iterasjonene
- Prosessen slutter når tapsfunskjonen (WCSS) er innenfor et visst tall





Bag of Words

Vi samler alle ord i en «bag», slik at vi får oversikt over hvilke ord som forekommer i et dokument, eventuelt hvor ofte de forekommer.

Rekkefølgen blir ikke tatt hensyn til

Visualisering:

https://www.youtube.com/watch?v=5I3Ei69I40s