Klassifikasjon & Evaluering

liljacs@uio.no

To oppgaver innen kategorisering

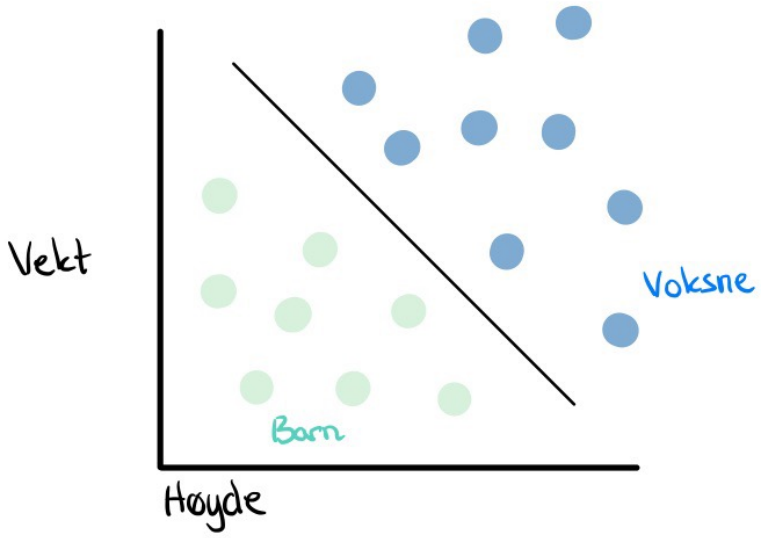
KLASSIFIKASJON

- *Veiledet* læring
- Krever *merket* treningsdata
- Tilordner forhåndsdefinerte klasser automatisk til nye instanser, gitt et sett av treningseksempler

KLYNGEANALYSE

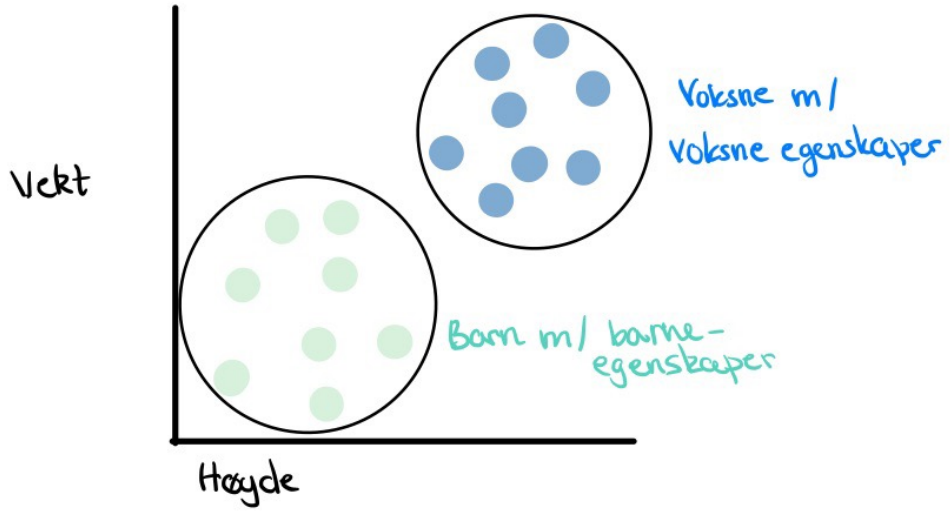
- *Ikke-veiledet* læring fra *umerket* treningsdata
- Grupperer automatisk like objekter sammen
- Ingen forhåndsdefinerte klasser
  - Spesifiserer kun likhetsmål

KLASSIFIKASJON



... Lærer av forhåndsdefinerte eksempler

KLYNGEANALYSE



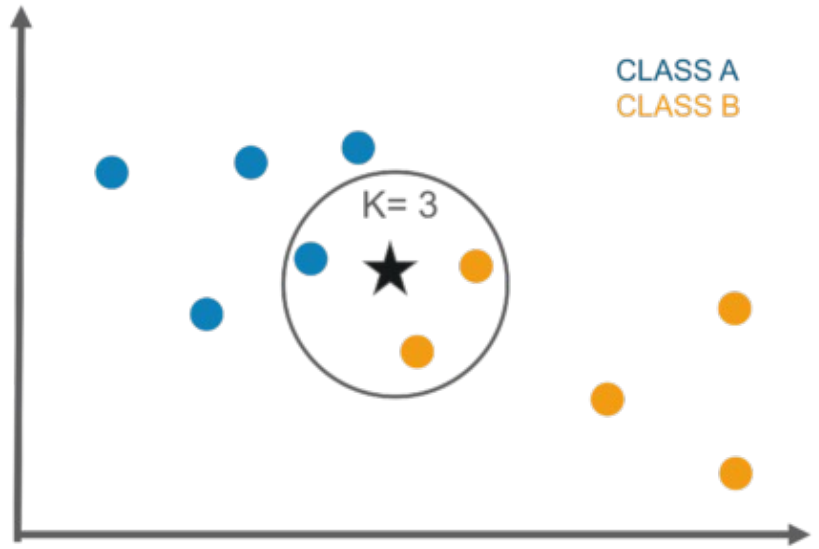
... Grupperer like objekter sammen

*k*NN

Multi-klasse-klassifiserer: Kan ha flere enn to klasser

Vi ser på de *k* nærmeste naboene til et objekt, og tilordner majoritetsklassen til objektet

Veiledet læring: vi forventer at et testobjekt i samme lokale region som et treningsobjekt, får tildelt samme klasse.



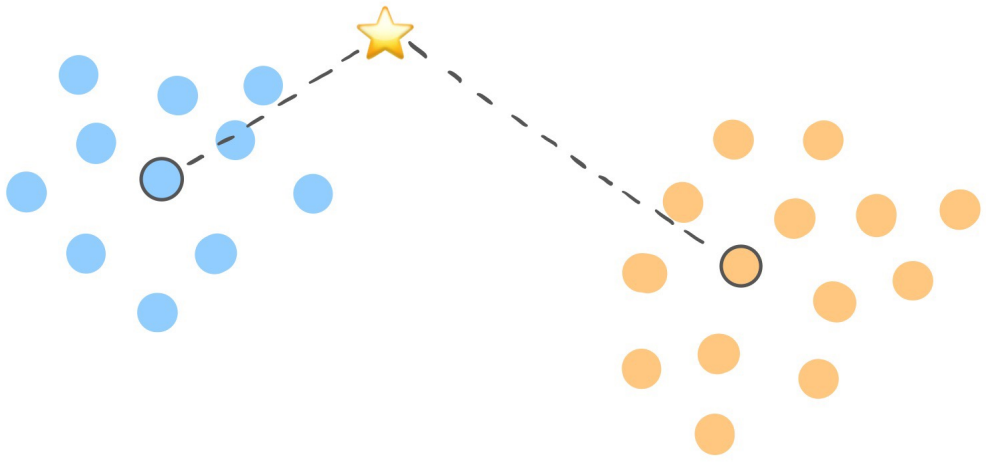
Her vil objektet ⭐ bli klassifisert som klasse B

Rocchio
«Nærmeste-centroide»

- Multi-klasse-klassifiserer
- Veiledet læring

Hver klasse representeres av dens centroide: «sentrum for gravitasjon»

Kalkuleres som gjennomsnittet av vektorene, tilhørende en klasse



Utregning av centroide

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i}$$

centroide — klasse — vektor

Gitt en klasse «musikk» med 3 treningsdokumenter:

X1 = (x11, x12, x13)
X2 = (x21, x22, x23)
X3 = (x31, x32, x33)

$$\mu Musikk = \frac{x11 + x21 + x31}{3}, \frac{x12 + x22 + x32}{3}, \frac{x13 + x23 + x33}{3}$$

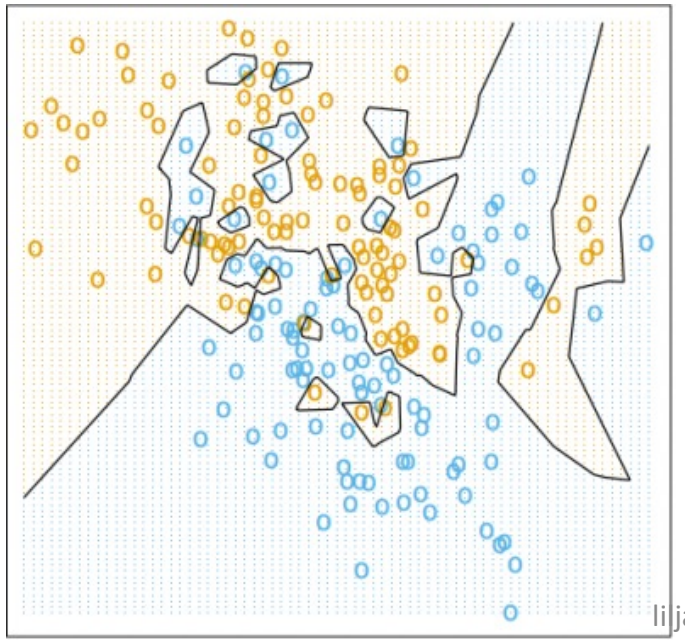
De tre verdiene vi får, utgjør den 3-dimensjonelle centroiden

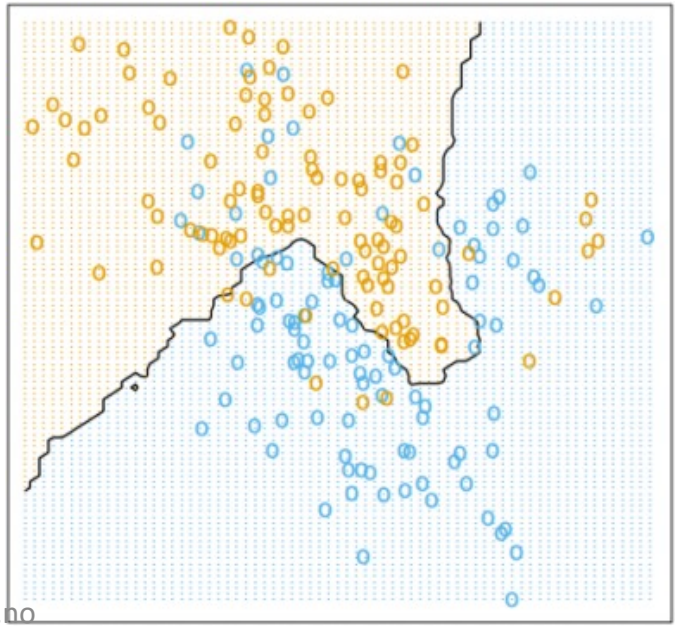
Hvordan velger vi hva *k* skal være?

For høy *k* kan føre til overgeneralisering/undertrening: klassifisereren er ikke nøyaktig nok på usette data

For lav *k* kan føre til overtrening: klassifisereren har lært seg treningsdataen for godt. Generaliserer dårlig

K=1





K=15

Evaluering

Hvordan måler vi hvor bra klassifikatoren vår gjør det?

$$\text{Accuracy} = \frac{TP+TN}{N}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-score: kombinerer precision og recall- «harmonic mean»

$$2 * \frac{precision * recall}{Precision + recall}$$