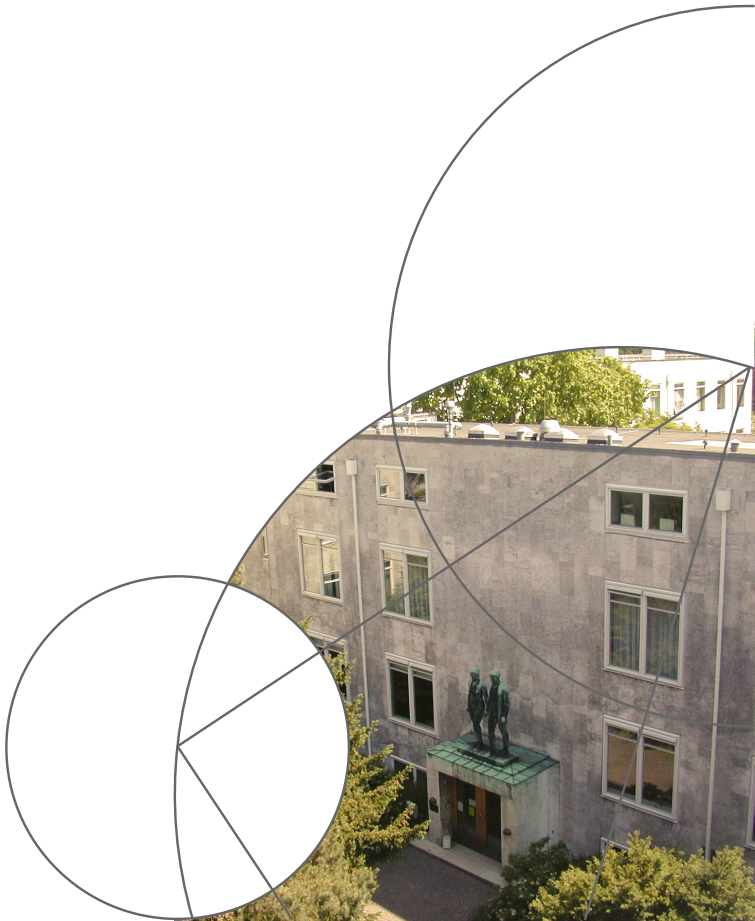
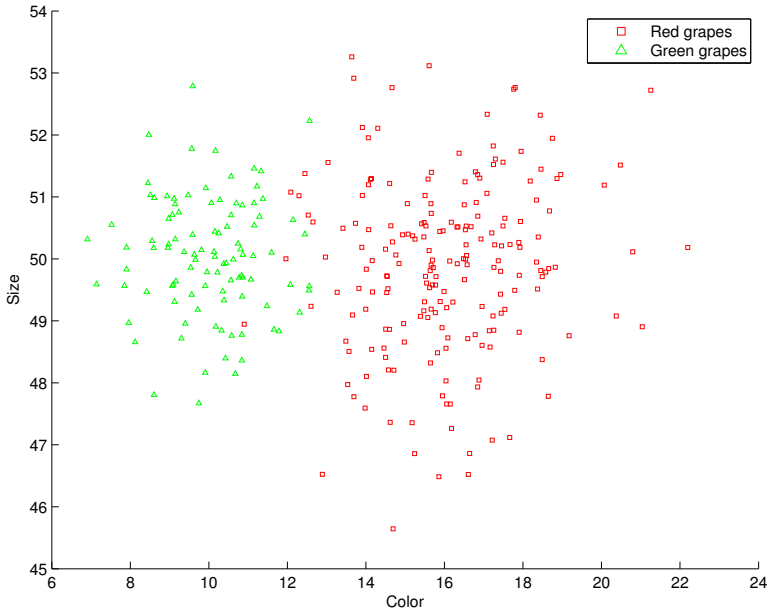
k-Nearest Neighbour

Peter Thougaard
& Nikolaj Overgaard Sørensen
Datalogisk Institut

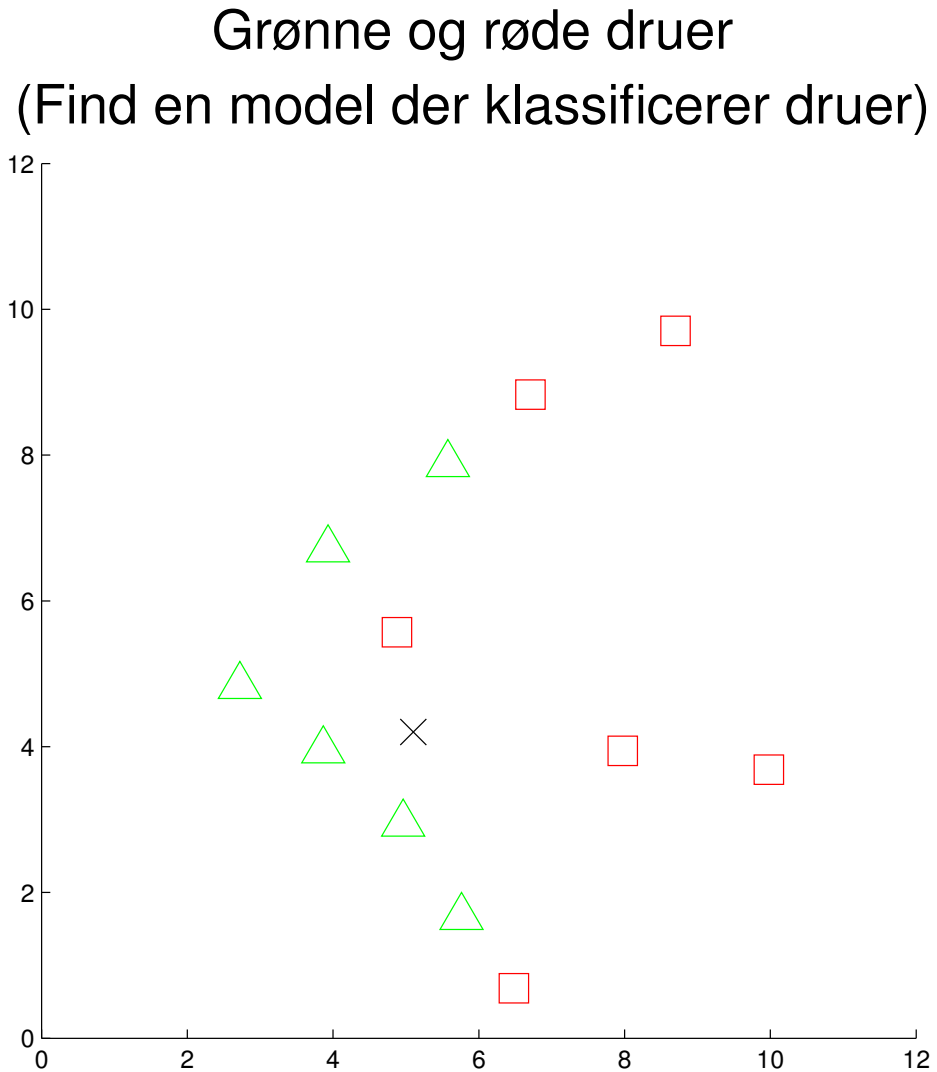


Vin-bonde i Frankrig

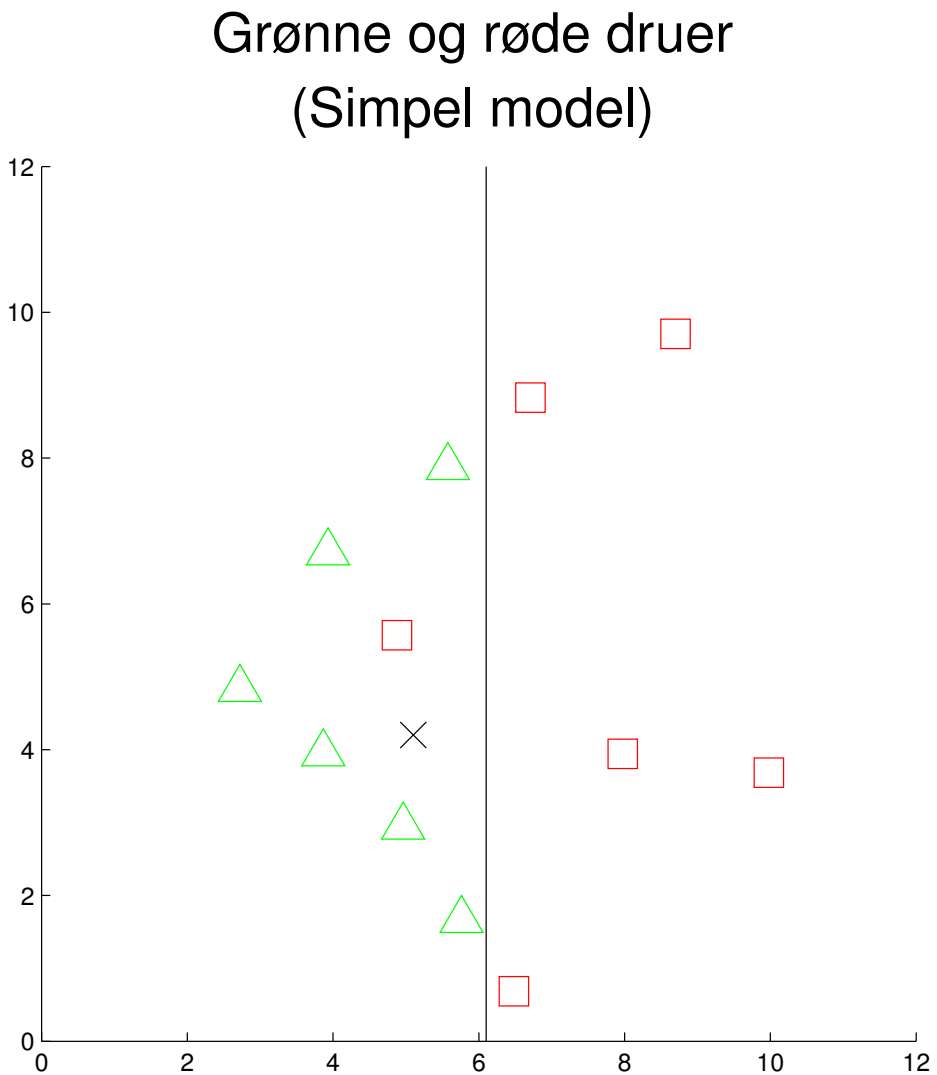
Grønne og røde druer



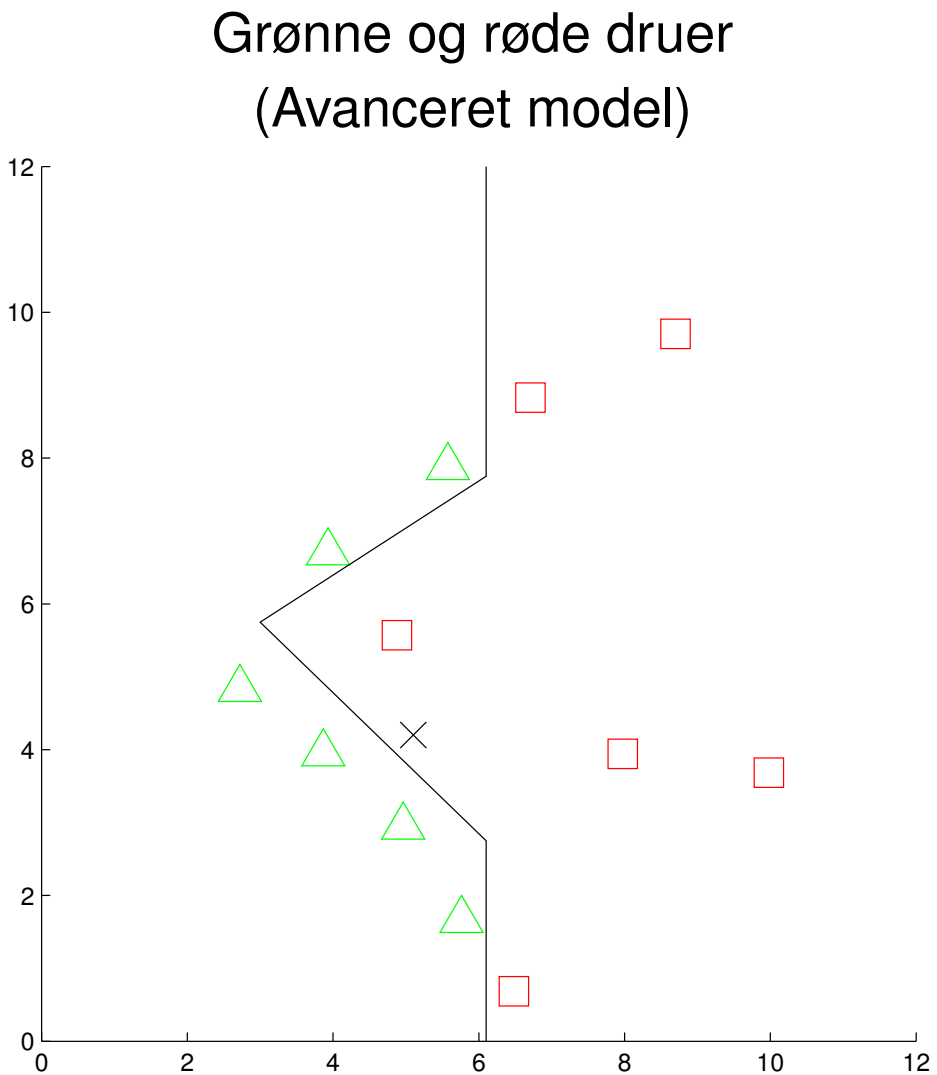
Vin-bonde i Frankrig



Vin-bonde i Frankrig



Vin-bonde i Frankrig

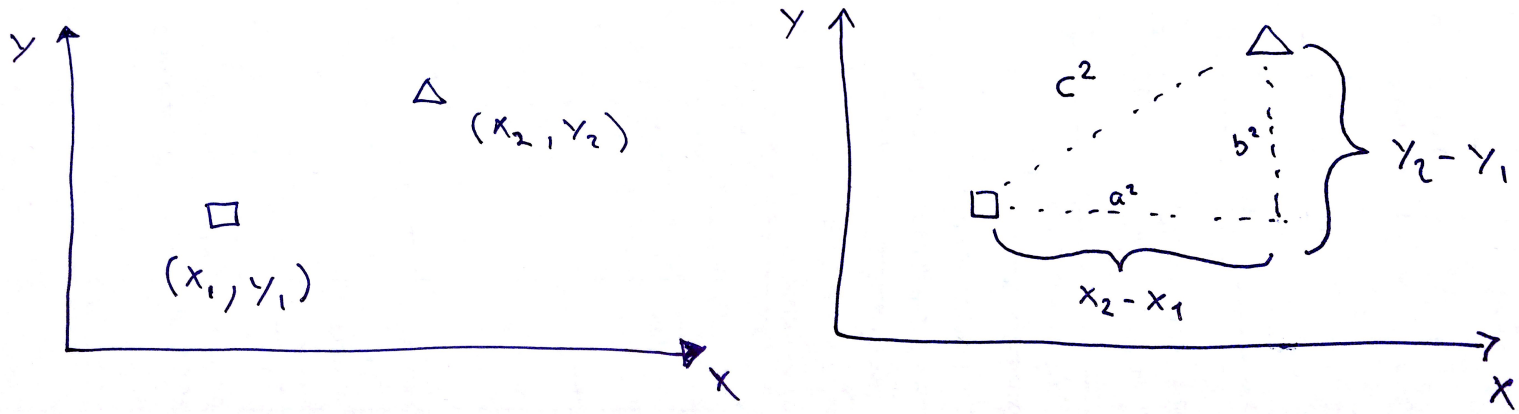


*k*NN - *k-nearest Neighbours*

*k*NN er forkortelse for *k-nearest Neighbours*

- de *k* nærmeste naboer findes ved mål af afstand mellem data-punkter
  (afstanden mellem deres koordinater)
- nyt punkt får samme kategori som flest naboer tilhører

*k*NN - Brug Pythagoras



Eksempel:

- Afstand, Pythagoras!
- $a^2 + b^2 = c^2 \Rightarrow c = \sqrt{a^2 + b^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- afstanden $d(x, y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Nearest neighbour - $k = 1$

Algoritme for $k = 1$

1. Vi ser på et nyt datapunkt P
2. Afstanden til hvert af kendte datapunkter $S_1, S_2, \ldots, S_n$ udregnes
3. Det nye punkt tildeles samme kategori som det punkt det er "tættest"på

Nearest neighbour - Vilkårligt k

Algoritme for k-NN

1. Vi ser på et nyt datapunkt P
2. Afstanden til hvert af kendte datapunkter $S_1, S_2, \ldots, S_n$ udregnes
3. De k "nærmeste"punkter identificeres
4. Det nye punkt tildeles samme kategori som deles af flest af disse k punkter

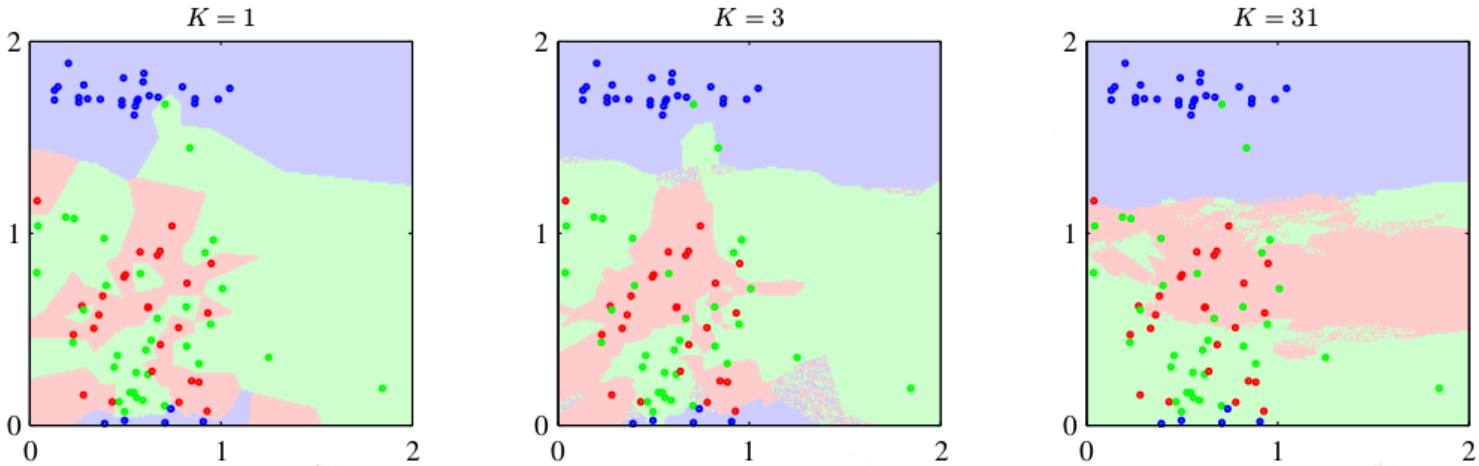
kNN - mere formel beskrivelse

Definition: Algoritme for k-Nearest Neighbor

1. Tag en mængde $S = \{S_1, S_2, \ldots, S_n\}$ af kendte data-punkter,
   med labels (rød firkant eller grøn trekant).
2. Tag en ukendt data-punkt P.
3. Udregn $d(P, S_i)$ for $i = 1, 2, \ldots, n$ og put dem i liste L.
4. Sorter liste L i faldende orden (mindste først).
5. Udvælg de k første punkter i L
6. Tildel P den label som optræder flest gange i L (uafgjort deles tilfældigt)
7. Gentag fra punkt 2 hvis der er flere ukendte P.

*k*NN - for valgfrit *k*

Betydning af *k*:



C. M. Bishop. Pattern Recognition and Machine Learning, Springer, 2006

- Prikkerne er vores datapunkter og deres farve angiver deres 'rigtige' kategori.
- Baggrundsfarven angiver den kategori nye punkter bliver tildelt.
- Skillelinjerne i baggrundsfarven viser hvor grænsen mellem tildelingen af kategorier går.

Lad os prøve det!