



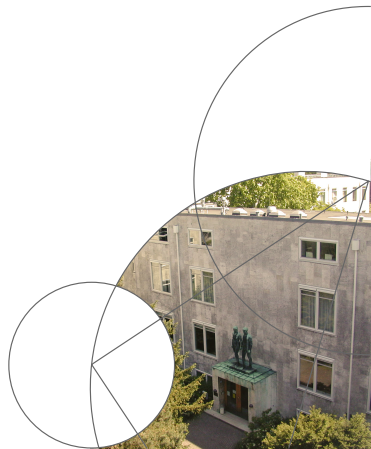
Det Naturvidenskabelige Fakultet



Studiepraktik dag 2

Datanalyse med Machine Learning

Arinbjörn Brandsson
Benjamin Rotendahl
Mathias Mortensen
Datalogisk Institut



- ① Hvad er Machine Learning
- ② Problemet
- ③ Algoritmen
- ④ Eksempel og algoritme-analyse



Hvad er Machine Learning

Problemstilling

Vi indsamler større og større mængder af data hele tiden, så meget at det har fået sit eget buzzword **Big Data**.



Hvad er Machine Learning

Problemstilling

Vi indsamler større og større mængder af data hele tiden, så meget at det har fået sit eget buzzword **Big Data**.

Vi mennesker kan ikke overskue så store mængder af data



Hvad er Machine Learning

Problemstilling

Vi indsamler stør og stør mængder af data hele tiden, så meget at det har fået sit eget buzzword **Big Data**.

Vi mennesker kan ikke overskue så store mængder af data

ML til undsætning!

Vi ønsker istedet at lave systemer sådan at computere kan finde de underliggende mønstre og bruge den viden/erfaring der ligger i data'en



Hvad er Machine Learning

Problemstilling

Vi indsamler stør og stør mængder af data hele tiden, så meget at det har fået sit eget buzzword **Big Data**.

Vi mennesker kan ikke overskue så store mængder af data

ML til undsætning!

Vi ønsker istedet at lave systemer sådan at computere kan finde de underliggende mønstre og bruge den viden/erfaring der ligger i data'en

Hvornår er ML godt?

- 1 Der eksisterer et mønster



Hvad er Machine Learning

Problemstilling

Vi indsamler stør og stør mængder af data hele tiden, så meget at det har fået sit eget buzzword **Big Data**.

Vi mennesker kan ikke overskue så store mængder af data

ML til undsætning!

Vi ønsker istedet at lave systemer sådan at computere kan finde de underliggende mønstre og bruge den viden/erfaring der ligger i data'en

Hvornår er ML godt?

- 1 Der eksisterer et mønster
- 2 Vi kan ikke finde en matematisk formel



Hvad er Machine Learning

Problemstilling

Vi indsamler stør og stør mængder af data hele tiden, så meget at det har fået sit eget buzzword **Big Data**.

Vi mennesker kan ikke overskue så store mængder af data

ML til undsætning!

Vi ønsker istedet at lave systemer sådan at computere kan finde de underliggende mønstre og bruge den viden/erfaring der ligger i data'en

Hvornår er ML godt?

- 1 Der eksisterer et mønster
- 2 Vi kan ikke finde en matematisk formel
- 3 Vi har data på problemet



Eksempel tid

Netflix udlovede en dusør på 6,5 millioner kroner til den der kunne forbedre deres anbefalings algoritme med 10%.



Eksempel tid

Netflix udlovede en dusør på 6,5 millioner kroner til den der kunne forbedre deres anbefalings algoritme med 10%.

Kan ML bruges?

- 1 Der eksisterer et mønster!



Eksempel tid

Netflix udlovede en dusør på 6,5 millioner kroner til den der kunne forbedre deres anbefalings algoritme med 10%.

Kan ML bruges?

- 1 Der eksisterer et mønster!
- 2 Vi kan ikke finde en formel for film



Eksempel tid

Netflix udlovede en dusør på 6,5 millioner kroner til den der kunne forbedre deres anbefalings algoritme med 10%.

Kan ML bruges?

- 1 Der eksisterer et mønster!
- 2 Vi kan ikke finde en formel for film
- 3 Der er massere af data til rådighed!



Eksempel tid

Netflix udlovede en dusør på 6,5 millioner kroner til den der kunne forbedre deres anbefalings algoritme med 10%.

Kan ML bruges?

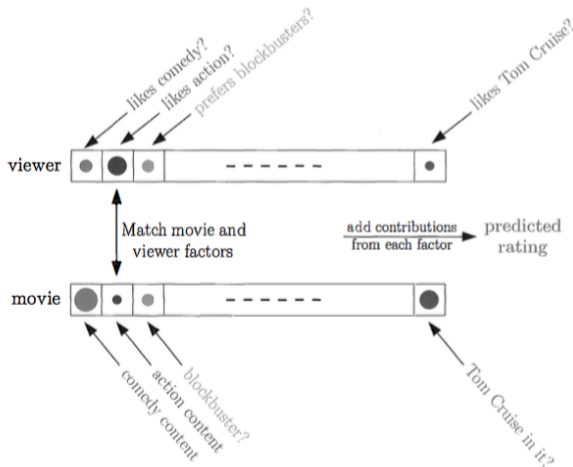
- 1 Der eksisterer et mønster!
- 2 Vi kan ikke finde en formel for film
- 3 Der er massere af data til rådighed!

ML vandt konkurrencen!



Hvordan vandt de?

Figure: Netflix vinderen



Dagens *Historie*

*Vi er blevet hyret af Danske Bank da de har hørt
at vi dataloger kan tjene dem en masse penge.*



Dagens *Historie*

Vi er blevet hyret af Danske Bank da de har hørt at vi dataloger kan tjene dem en masse penge.

Problemstilling

Vi skal lave et system der givet data om en kunde og et beløb kan bestemme om det er en god forretning at låne dem de penge.



Dagens *Historie*

Vi er blevet hyret af Danske Bank da de har hørt at vi dataloger kan tjene dem en masse penge.

Problemstilling

Vi skal lave et system der givet data om en kunde og et beløb kan bestemme om det er en god forretning at låne dem de penge.

Hmm, det var da et ret generelt problem ...



Dagens *Historie*

Vi er blevet hyret af Danske Bank da de har hørt at vi dataloger kan tjene dem en masse penge.

Problemstilling

Vi skal lave et system der givet data om en kunde og et beløb kan bestemme om det er en god forretning at låne dem de penge.

Hmm, det var da et ret generelt problem ...

Problemstilling

Vi skal lave et system der givet data om en **patient** og en **mængde af Chemo** kan bestemme om det er en god behandling.



Dagens *Historie*

Vi er blevet hyret af Danske Bank da de har hørt at vi dataloger kan tjene dem en masse penge.

Problemstilling

Vi skal lave et system der givet data om en kunde og et beløb kan bestemme om det er en god forretning at låne dem de penge.

Hmm, det var da et ret generelt problem ...

Problemstilling

Vi skal lave et system der givet data om en **patient** og en **mængde af Chemo** kan bestemme om det er en god behandling.

Vi koder for kapitalen!



- 1 Hvad er Machine Learning
- 2 Problemet
- 3 Algoritmen
- 4 Eksempel og algoritme-analyse



Håndtering af input og output

Input

alder	27
køn	1
Årlig Løn	50.000
Bopæl	2300
Gæld	40.000
⋮	⋮
Ønsket lån	3.000.000

Output

Tjente vi penge? ja



Håndtering af input og output

Input

alder	27
køn	1
Årlig Løn	50.000
Bopæl	2300
Gæld	40.000
⋮	⋮
Ønsket lån	3.000.000



Data vektor

$$\begin{pmatrix} 27 \\ 1 \\ 50.000 \\ 2300 \\ 40.000 \\ \vdots \\ 3.000.000 \end{pmatrix}$$

Output

Tjente vi penge? ja

Output

1



Formalisering

Termer

Input En vektor (Lån ansøgning)



Formalisering

Termer

Input En vektor (Lån ansøgning)

Output 1 eller -1 (God eller dårlig forretning)



Formalisering

Termer

Input En vektor (Lån ansøgning)

Output 1 eller -1 (God eller dårlig forretning)

Læringsmål $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$



Formalisering

Termer

Input En vektor (Lån ansøgning)

Output 1 eller -1 (God eller dårlig forretning)

Læringsmål $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$

Data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ (Hvad vi lærer fra)



Formalisering

Termer

Input En vektor (Lån ansøgning)

Output 1 eller -1 (God eller dårlig forretning)

Læringsmål $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$

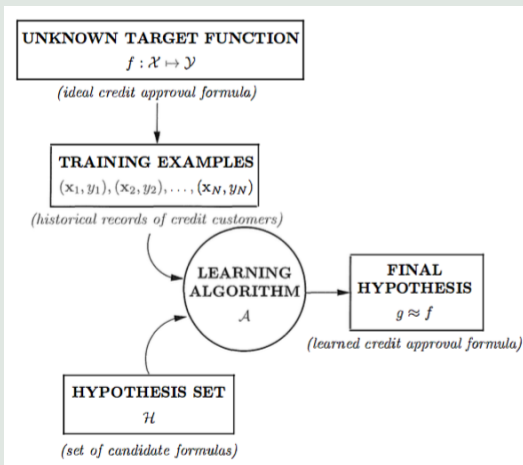
Data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ (Hvad vi lærer fra)

Hypotese $g : \mathcal{X} \rightarrow \mathcal{Y}$ (Vores systems "Hjerne")



Visuel Formalisering

Figure: Visuelt læringsdiagram



- 1 Hvad er Machine Learning
- 2 Problemet
- 3 Algoritmen**
- 4 Eksempel og algoritme-analyse



Valget af lærings-algoritmen

Perceptron

Den laver et *hyperplan* der adskiler data'en og finder en opdeling der giver en **lav fejl**.



Valget af lærings-algoritmen

Perceptron

Den laver et *hyperplan* der adskiler data'en og finder en opdeling der giver en **lav fejl**.

Tænk på den som en form for lineær regression på steroider

$$y = ax + b$$



Valget af lærings-algoritmen

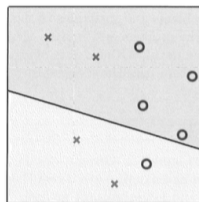
Perceptron

Den laver et *hyperplan* der adskiler data'en og finder en opdeling der giver en **lav fejl**.

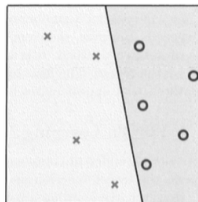
Tænk på den som en form for lineær regression på steroider

$$y = ax + b$$

Eksempel på algoritmen



(a) Misclassified data



(b) Perfectly classified data



Algoritmen i ord

Hvordan virker den?

Vi har en masse vektorer v_1, v_2, \dots, v_n og en liste af svar y_1, y_2, \dots, y_n .



Algoritmen i ord

Hvordan virker den?

Vi har en masse vektorer v_1, v_2, \dots, v_n og en liste af svar y_1, y_2, \dots, y_n .

Vi lader w være vores “vægt-vektor”.



Algoritmen i ord

Hvordan virker den?

Vi har en masse vektorer v_1, v_2, \dots, v_n og en liste af svar y_1, y_2, \dots, y_n .

Vi lader w være vores “vægt-vektor”.

Godkend lån hvis : $\sum_{i=1}^d w_i x_i > b$

Afvis lån hvis : $\sum_{i=1}^d w_i x_i < b$



Algoritmen i ord

Hvordan virker den?

Vi har en masse vektorer v_1, v_2, \dots, v_n og en liste af svar y_1, y_2, \dots, y_n .

Vi lader w være vores “vægt-vektor”.

Godkend lån hvis : $\sum_{i=1}^d w_i x_i > b$

Afvis lån hvis : $\sum_{i=1}^d w_i x_i < b$

Vores hypotese bliver så

$$h(x) = \text{fortegn} \left(\sum_{i=0}^d w_i x_i \right)$$



Algoritmen i ord

Hvordan virker den?

Vi har en masse vektorer v_1, v_2, \dots, v_n og en liste af svar y_1, y_2, \dots, y_n .

Vi lader w være vores “vægt-vektor”.

Godkend lån hvis : $\sum_{i=1}^d w_i x_i > b$

Afvis lån hvis : $\sum_{i=1}^d w_i x_i < b$

Vores hypotese bliver så

$$h(x) = \text{fortegn} \left(\sum_{i=0}^d w_i x_i \right)$$

Men hvordan bestemmer vi w ?



Hvordan den lærer

Hvordan w bestemmes



Hvordan den lærer

Hvordan w bestemmes

$w =$ vælg tilfældige tal



Hvordan den lærer

Hvordan w bestemmes

$w =$ vælg tilfældige tal

Vi forbederer w hver gang!



Hvordan den lærer

Hvordan w bestemmes

$w =$ vælg tilfældige tal

Vi forbederer w hver gang!

Hvis x' er på den forkerte side af w så lærer den
“erfaringen” ved formlen



Hvordan den lærer

Hvordan w bestemmes

$w =$ vælg tilfældige tal

Vi forbederer w hver gang!

Hvis x' er på den forkerte side af w så lærer den “erfaringen” ved formlen

$$w_{ny} = w + y'x'$$



Hvordan den lærer

Hvordan w bestemmes

$w =$ vælg tilfældige tal

Vi forbederer w hver gang!

Hvis x' er på den forkerte side af w så lærer den “erfaringen” ved formlen

$$w_{ny} = w + y'x'$$

Forsæt med at lære indtil du ikke kan lære mere.



Perceptron algoritme

Pseudocode

Algorithm 1

Input: datasæt $X = [(x_1, y_1), \dots, (x_n, y_n)]$

Output: Hypotesen w .

w = Tilfældige tal

$\text{misCat} = (1, 1)$

while $\text{misCat} \neq (0, 0)$ **do**

$\text{misCat} = (0, 0)$

for (x_i, y_i) in X **do**

if $\text{sign}(w^T x_i) \neq y_i$ **then**

$\text{misCat} = (x_i, y_i)$

$w = w + y_i x_i$

end if

end for

end while

return w

- ① Hvad er Machine Learning
- ② Problemet
- ③ Algoritmen
- ④ Eksempel og algoritme-analyse



Vi prøver at køre den!

Eksempel i MatLab



Vi prøver at køre den!

Eksempel i MatLab

Analyse

Nogen der kan gætte køretiden?



Vi prøver at køre den!

Eksempel i MatLab

Analyse

Nogen der kan gætte køretiden?

$$O(2^{(n+1)\log(n+1)}(n+1)^2)$$

Tjener vi så nogle penge eller redder nogle liv?



Vi prøver at køre den!

Eksempel i MatLab

Analyse

Nogen der kan gætte køretiden?

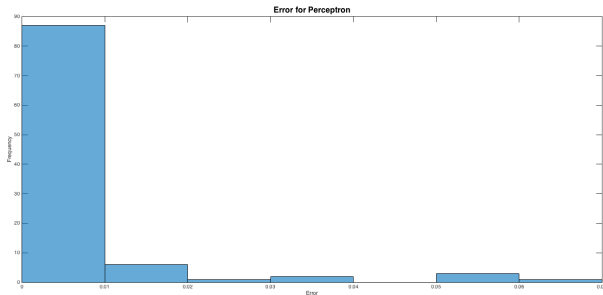
$$O(2^{(n+1)\log(n+1)}(n+1)^2)$$

Tjener vi så nogle penge eller redder nogle liv? Det kan jeg ikke svare på ☹



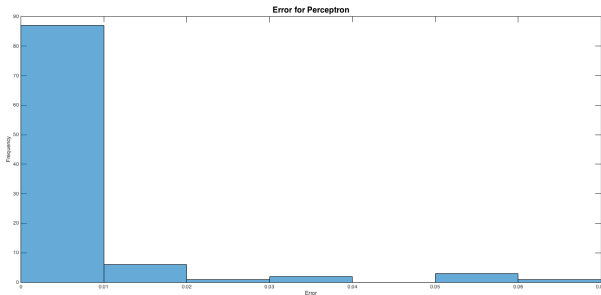
Afslutning

Fejl i min simulering



Afslutning

Fejl i min simulering

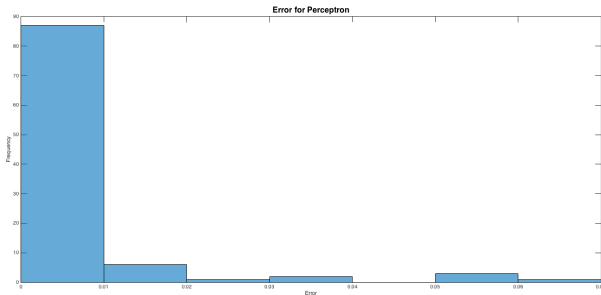


Spørgsmål?



Afslutning

Fejl i min simulering



Spørgsmål?
Har jeg tid til mere?

