IT FAGENE I DE GYMNASIALE UDDANNELSER

Machine learning, datamining og big data

Benjamin Rotendahl

April 17, 2016

Hvem er det?

- Kæmpe nørd.

Hvem er det?

- Kæmpe nørd.
- Datalogistuderende ved Københavns Univsersitet.

Hvem er det?

- Kæmpe nørd.
- Datalogistuderende ved Københavns Univsersitet.
- Frivilig/studentermedhjælper/bestyrelsesmedlem i Coding Pirates.

Hvem er det?

- Kæmpe nørd.
- Datalogistuderende ved Københavns Univsersitet.
- Frivilig/studentermedhjælper/bestyrelsesmedlem i Coding Pirates.
- Medlem af Datalogisk Instituts gymnasietjeneste.

PROGRAMMET FOR IDAG

Intro til Machine Learning

Forklaring af de overorddne ideer og tanker bag machine learning

PROGRAMMET FOR IDAG

Intro til Machine Learning

Forklaring af de overorddne ideer og tanker bag machine learning

Perceptron algoritmen

En simpel machine learning algortime bygget på vektorregning.

PROGRAMMET FOR IDAG

Intro til Machine Learning

Forklaring af de overorddne ideer og tanker bag machine learning

Perceptron algoritmen

En simpel machine learning algortime bygget på vektorregning.

Fremvising og øvelser i iPythone

Interaktive øvelser i det fremragende iPython.

PROGRAMMET FOR IDAG

Intro til Machine Learning

Forklaring af de overorddne ideer og tanker bag machine learning

Perceptron algoritmen

En simpel machine learning algortime bygget på vektorregning.

Fremvising og øvelser i iPythone

Interaktive øvelser i det fremragende iPython.

Afrunding og spørgsmål

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.

Løsning

Finde en måde at få computere til at finde de underliggende mønstre og bruge den viden/erfaring der ligger i data'en.

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.

Løsning

Finde en måde at få computere til at 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 større og større mængder af data hele tiden, så meget at det har fået sit eget buzzword Big Data.

Løsning

Finde en måde at få computere til at 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 større og større mængder af data hele tiden, så meget at det har fået sit eget buzzword Big Data.

Løsning

Finde en måde at få computere til at 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

DAGENS ØVELSE

Vi er blevet hyret af et hospital da de har hørt at vi IT-folk kan hjælpe deres patienter.

DAGENS ØVELSE

Vi er blevet hyret af et hospital da de har hørt at vi IT-folk kan hjælpe deres patienter.

Problemstilling

Vi skal lave et system der, givet data om en patient, kan bestemme om deres svulst er godartet eller ondartet.

DAGENS ØVELSE

Vi er blevet hyret af et hospital da de har hørt at vi IT-folk kan hjælpe deres patienter.

Problemstilling

Vi skal lave et system der, givet data om en patient, kan bestemme om deres svulst er godartet eller ondartet.

Hmm, det var da et ret generelt problem …

DAGENS ØVELSE

Vi er blevet hyret af et hospital da de har hørt at vi IT-folk kan hjælpe deres patienter.

Problemstilling

Vi skal lave et system der, givet data om en patient, kan bestemme om deres svulst er godartet eller ondartet.

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

Problemstilling

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

DAGENS ØVELSE

Vi er blevet hyret af et hospital da de har hørt at vi IT-folk kan hjælpe deres patienter.

Problemstilling

Vi skal lave et system der, givet data om en patient, kan bestemme om deres svulst er godartet eller ondartet.

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

Problemstilling

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

Problemet kaldes klassificering, gode løsninger kan have stor indflydelse inden for mange felter.

Kan Machine Learning bruges?

Der ekisterer et mønster

Kræftsvulsters udvikling er ikke tilfældig.

EKSEMPEL

Kan Machine Learning bruges?

Der ekisterer et mønster

Kræftsvulsters udvikling er ikke tilfældig.

Svært at finde en formel

Der er støj i data'en og det er ikke konsekvent.

EKSEMPEL

Kan Machine Learning bruges?

Der ekisterer et mønster

Kræftsvulsters udvikling er ikke tilfældig.

Svært at finde en formel

Der er støj i data'en og det er ikke konsekvent.

Vi har data på problemet

Data fra Washington state hospital.

FORMALISERING

Termer

Input: En vektor (patient data)

Termer

Input: En vektor (patient data)

Output: 1 eller -1 (ondartet eller godartet)

FORMALISERING

Termer

Input: En vektor (patient data)

Output: 1 eller -1 (ondartet eller godartet)

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

FORMALISERING

Termer

Input: En vektor (patient data)

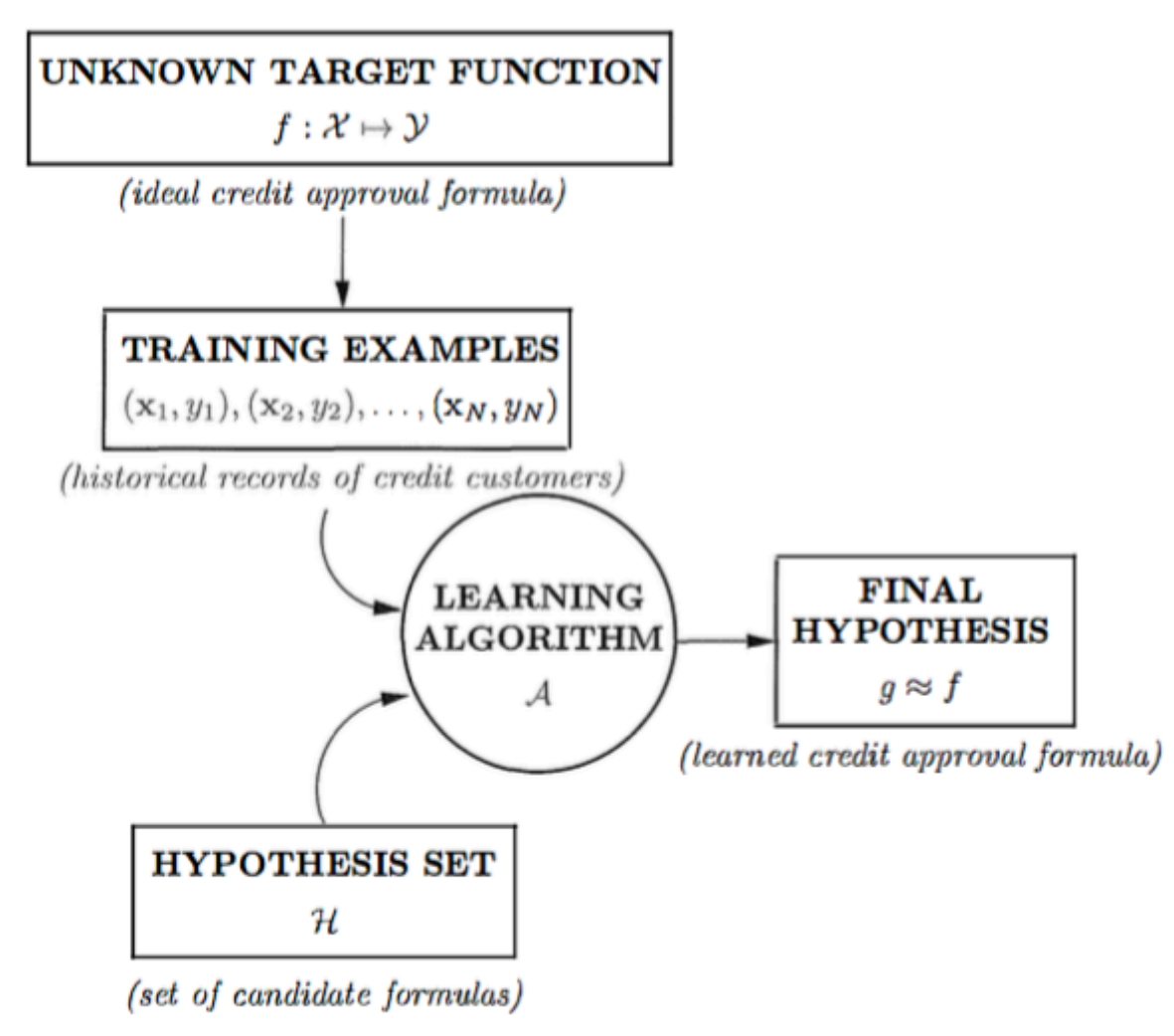
Output: 1 eller -1 (ondartet eller godartet)

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

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

VISUEL FORMALISERING

Figure 1: Visuelt læringsdiagram



Et kig på vores data?

Input

Threshold	1
Clump Thickness	7
Uniformity of Cell Size	1
Uniformity of Cell Shape	4
Epithelial Cell Size	2
Bare Nuclei	3
Bland Chromatin	8
Normal Nucleoli	10
Mitoses	3

Output

Ondartet eller godartet

ET KIG PÅ VORES DATA?

Input

Threshold	1
Clump Thickness	7
Uniformity of Cell Size	1
Uniformity of Cell Shape	4
Epithelial Cell Size	2
Bare Nuclei	3
Bland Chromatin	8
Normal Nucleoli	10
Mitoses	3

Output

Ondartet eller godartet

Data vektor

$$\begin{pmatrix} 1 \\ 7 \\ 1 \\ 4 \\ 2 \\ 3 \\ 8 \\ 10 \\ 3 \end{pmatrix}$$

Output

1

PERCEPTRON ALGORITMEN

VALGET AF LÆRINGS-ALGORITMEN

Perceptron

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

VALGET AF LÆRINGS-ALGORITMEN

Perceptron

Den laver et hyperplan der adskiller 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

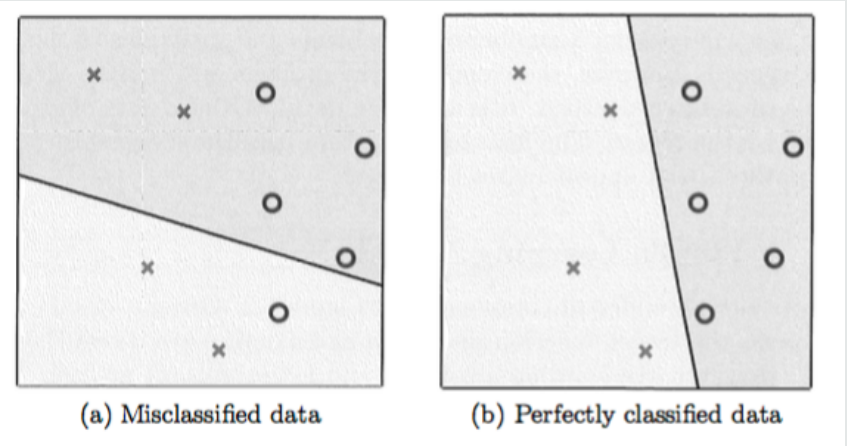
Perceptron

Den laver et hyperplan der adskiller 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

Hvordan virker den?

Vi har en masse vektorer $x_1, x_2, \ldots, x_n$ og en liste af svar $y_1, y_2, \ldots, y_n$.

Hvordan virker den?

Vi har en masse vektorer $x_1, x_2, \ldots, x_n$ og en liste af svar $y_1, y_2, \ldots, y_n$.
Vi lader w være vores "hjerne-vektor".

Hvordan virker den?

Vi har en masse vektorer $x_1, x_2, \ldots, x_n$ og en liste af svar $y_1, y_2, \ldots, y_n$.

Vi lader w være vores "hjerne-vektor".

$$\text{Godartet svulst} : \sum_{i=1}^{d} w_i x_i > b$$

$$\text{Ondartet svulst} : \sum_{i=1}^{d} w_i x_i < b$$

Hvordan virker den?

Vi har en masse vektorer $x_1, x_2, \ldots, x_n$ og en liste af svar $y_1, y_2, \ldots, y_n$.

Vi lader w være vores "hjerne-vektor".

$$\text{Godartet svulst} : \sum_{i=1}^{d} w_i x_i > b$$

$$\text{Ondartet svulst} : \sum_{i=1}^{d} w_i x_i < b$$

Vores hypotese bliver så

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

Hvordan virker den?

Vi har en masse vektorer $x_1, x_2, \ldots, x_n$ og en liste af svar $y_1, y_2, \ldots, y_n$.

Vi lader w være vores "hjerne-vektor".

$$\text{Godartet svulst} : \sum_{i=1}^{d} w_i x_i > b$$

$$\text{Ondartet svulst} : \sum_{i=1}^{d} w_i x_i < b$$

Vores hypotese bliver så

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

Men hvordan bestemmer vi w?

Hvordan w bestemmes

$$\mathrm{w} = \text{ vælg tilfældige tal}$$

Hvordan w bestemmes

$$\mathrm{w} = \text{ vælg tilfældige tal}$$

Vi forbedrer w hver gang

Hvordan w bestemmes

$$w = \text{ vælg tilfældige tal}$$

Vi forbedrer w hver gang

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

Hvordan w bestemmes

$$w = \text{ vælg tilfældige tal}$$

Vi forbedrer w hver gang

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

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

Hvordan w bestemmes

$$w = \text{ vælg tilfældige tal}$$

Vi forbedrer 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 forbedre så længe så muligt.

PERCEPTRON ALGORITME

Pseudocode

w = Tilfældige tal
isLearning = True
while isLearning do
    isLearning = False
    for (x_i, y_i) in X do
        if $\text{sign}(w^T x_i) \neq y_i$ then
            isLearning = True
            $w = w + y_i x_i$
        end if
    end for
end while
return w

EKSEMPEL OG ALGORITME-ANALYSE

Hvad er Machine Learning

Perceptron algoritmen

Eksempel og algoritme-analyse

VI PRØVER AT KØRE DEN!

Analyse

VI PRØVER AT KØRE DEN!

Analyse

Redder vi så nogle liv?

VI PRØVER AT KØRE DEN!

Analyse

Redder vi så nogle liv? Lad os kode det og se hvor god den er!

Hvor god er den?

I opgaverne kigger i kun på 25 eksempler! og tester på 75 patienter

Hvor god er den?

I opgaverne kigger i kun på 25 eksempler! og tester på 75 patienter

I kan forvente at den har ret på cirka $60-70\%$ af patienterne!.

Hvor god er den?

I opgaverne kigger i kun på 25 eksempler! og tester på 75 patienter

I kan forvente at den har ret på cirka $60-70\%$ af patienterne!.

Kører man den istedet med 500 eksempler og tester på 180.

Hvor god er den?

I opgaverne kigger i kun på 25 eksempler! og tester på 75 patienter

I kan forvente at den har ret på cirka $60-70\%$ af patienterne!.

Kører man den istedet med 500 eksempler og tester på 180. Rammer den rigtigt 181 gange og forkert 2 gange. Det betyder at den har en succes rate på $98,9\%$!

Hvor god er den?

I opgaverne kigger i kun på 25 eksempler! og tester på 75 patienter

I kan forvente at den har ret på cirka $60 - 70\%$ af patienterne!.

Kører man den istedet med 500 eksempler og tester på 180. Rammer den rigtigt 181 gange og forkert 2 gange. Det betyder at den har en succes rate på $98,9\%$!

Spørgsmål?