IT FAGENE I DE GYMNASIALE UDDANNELSER

Machine learning, datamining og big data

Benjamin Rotendahl

April 17, 2016

Hvem er jeg?

- Datalogistuderende ved Københavns Universitet.
- Frivillig/studentermedhjælper/bestyrelsesmedlem i Coding Pirates.
- Medlem af Datalogisk Instituts ved Københavns universitets gymnasietjeneste.
- Slides kan hentes her “rotendahl.dk/slides.pdf”

Materiale

Udrag fra en fagpakke der er udviklet for gymnasietjenesten ved Datalogisk Institut Københavns Universitet.
Link : "https://github.com/Rotendahl/Gymnasie-tjeneste"

Undervisningsmetode

- Materialet er svært.
- Learn by doing.
- Smagsprøve.

PROGRAMMET FOR I DAG

Intro til Machine Learning

Forklaring af de overordnede ideer og tanker bag machine learning

Perceptron algoritmen

En simpel machine learning algoritme bygget på vektorregning.

Fremvisning og øvelser i iPython

Interaktive øvelser i værktøjet iPython.

Afrunding og spørgsmål

HVAD ER MACHINE LEARNING

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.

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.

FORMALISERING

Termer

Input: En vektor (patient data)

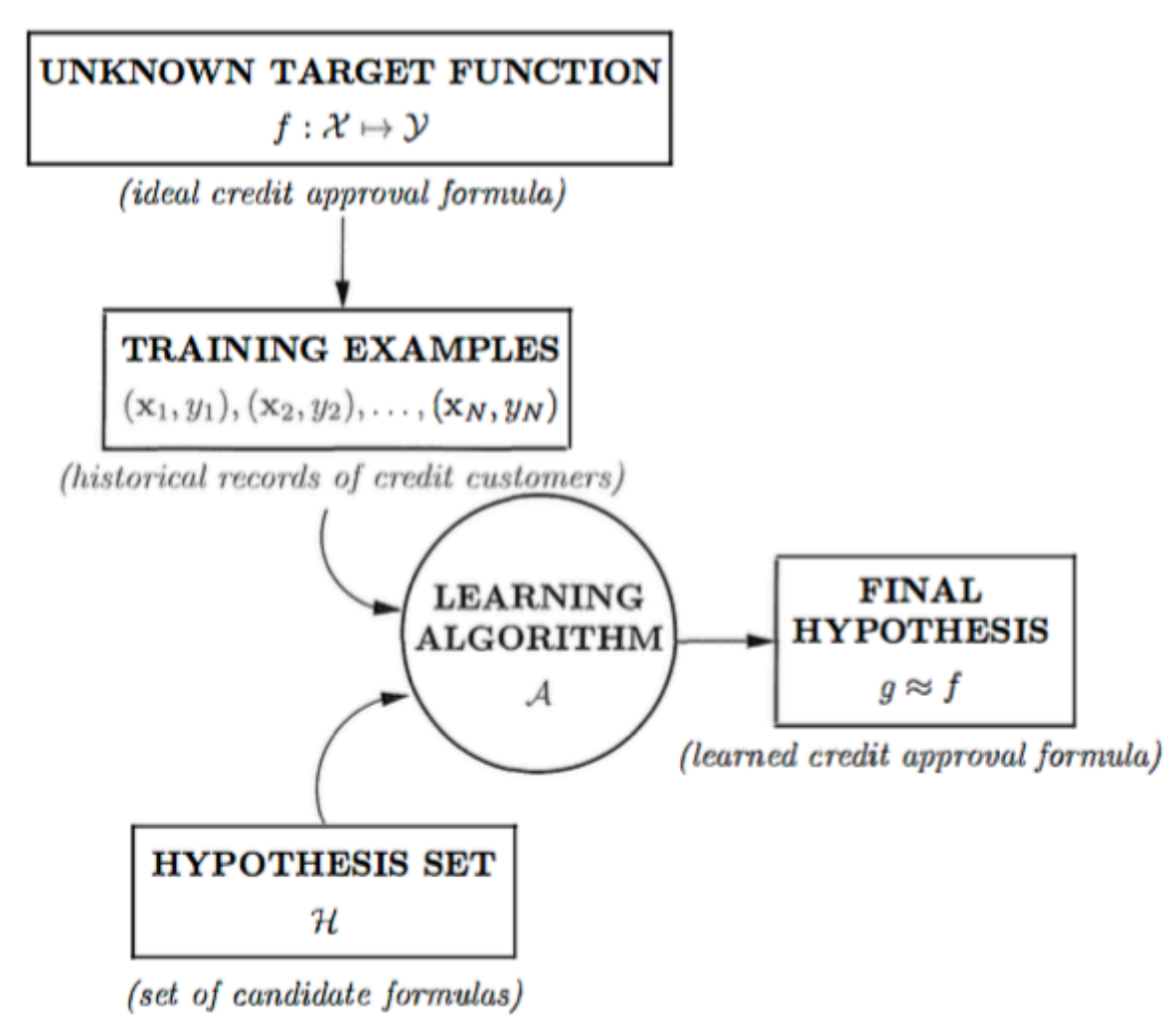
Output: 1 eller -1 (ondartet eller godartet)

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

Data vektor

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

Output

1

PERCEPTRON ALGORITMEN

VALGET AF LÆRINGS-ALGORITMEN

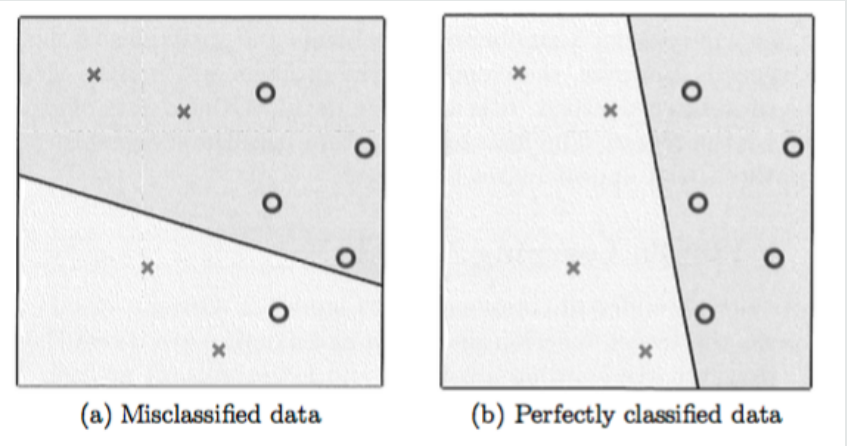
Perceptron

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

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 $y_1, y_2, \ldots, y_n$.
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

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

Men hvordan bestemmer vi w?

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 $sign(w^T x_i) \neq y_i$ then
            isLearning = True
            $w = w + y_i x_i$
        end if
    end for
end while
return w

IPYTHON OG ØVELSER

ØVELSER

iPython

Online python fortolker, med mulighed for nemt og hurtigt at lave opgaver til studerende.
Gå ind på “rotendahl.dk/vejle”
kode: DIKU

AFSLUTNING

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 i stedet med 500 eksempler og tester på 180. Rammer den rigtigt 178 gange og forkert 2 gange. Det betyder at den har en succes rate på $98,9\%$!

EMNER INDEN FOR MACHINE LEARNING

Matematikken bag

Hvordan sikrer vi at den faktisk kan sige noget om virkeligheden?

"https://work.caltech.edu/telecourse.html"

Etiske spørgsmål

Fordomme i data'en bliver lært, skal det være sådan?

Hvad betyder det at lære? "rotendahl.dk/MLexcerpt"

Automatisering

Skal vi have grænser for hvilke jobs de må tage?

"https://www.youtube.com/watch?v=7Pq-S557XQU"

Guide til iPython

Hvordan sætter man det op og hvor kan det ellers bruges til? "http://rotendahl.dk/iPython"

SPØRGSMÅL

Links

Benjamin@Rotendahl.dk

rotendahl.dk/slides.pdf

Tak for opmærksomheden

Spørgsmål?