KU KONKUK UNIVERSITY

KAIST

Introduction to Natural Language Processing

건국대학교 컴퓨터공학부 /
KAIST 전산학부 (겸직)

김학수

자연어처리연구실

1967

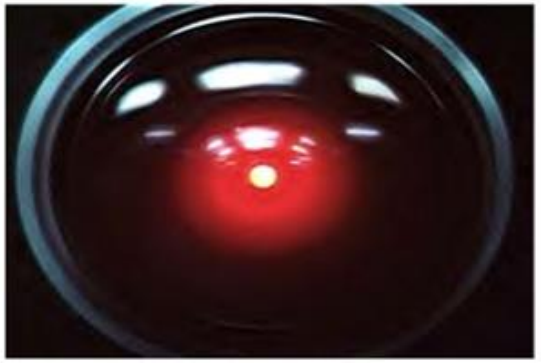
Stanley Kubrick,
Filmmaker
1928 - 1999

Arthur C. Clarke,
Author, futurist,
1917 – 2008


CAUTION
WEIGHTLESS
CONDITION


M G M DVD
A STANLEY KUBRICK PRODUCTION
2001:
a space odyssey

HAL

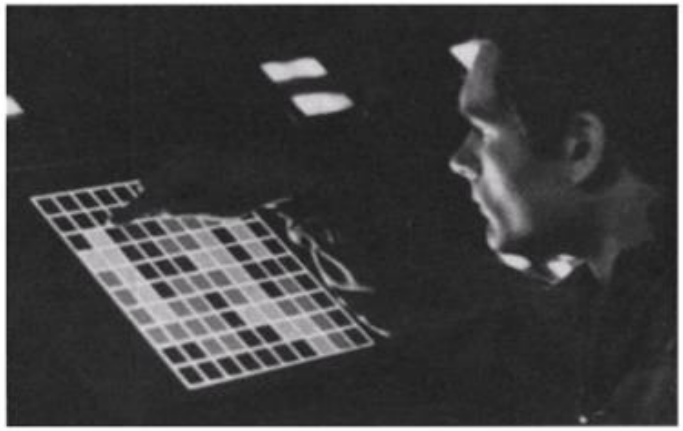


HAL's Capabilities

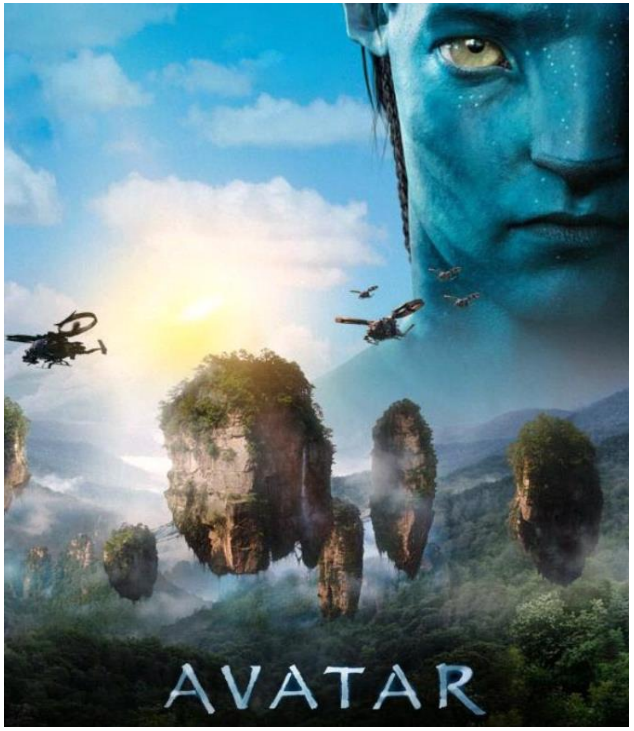
- Display graphics
- Play chess
- *Natural language understanding and generation*
- Vision
- Planning
- Learning
- …

Graphics

HAL



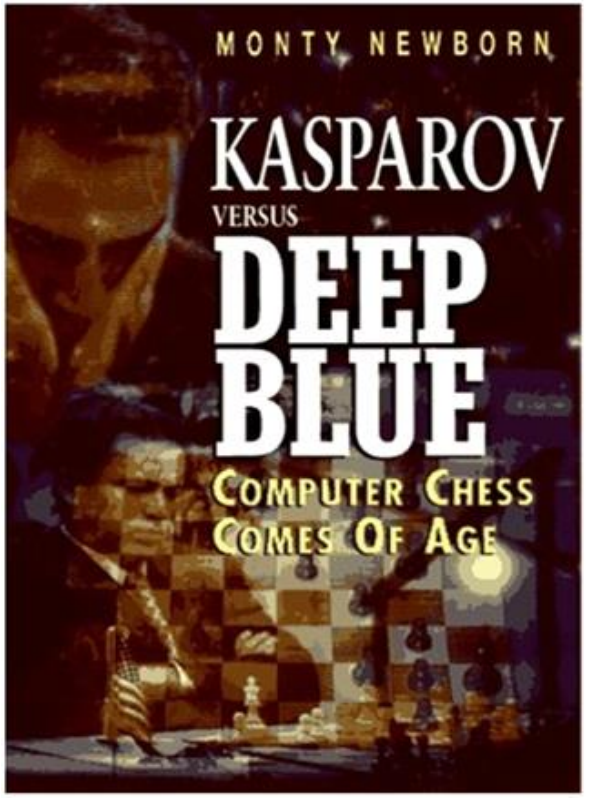
A few years ago



Chess

HAL

Now



Natural Language Understanding

HAL

David Bowman:

Open the pod bay doors, Hal.

HAL:

I'm sorry, Dave, I'm afraid I can't do that.

David Bowman:

What are you talking about, Hal?

…

HAL:

I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

Now

?

Many useful tools, but none that come even close to HAL's ability to communicate in natural language.

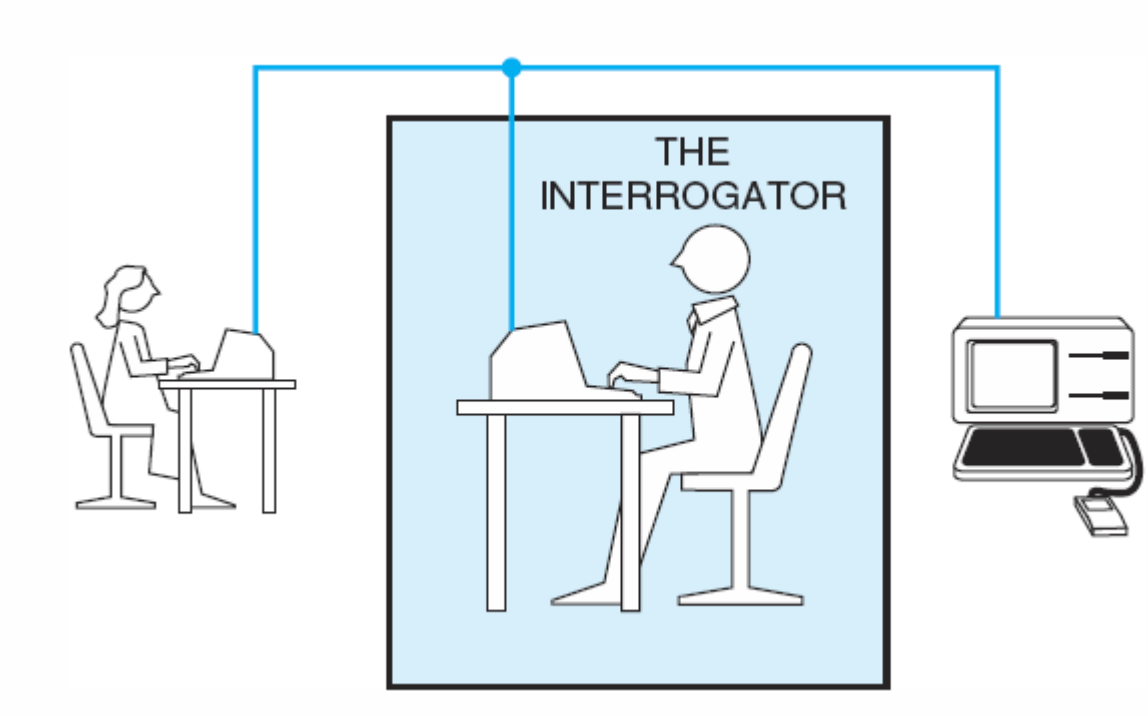
AI에 대한 정의들

- 지능적 행동의 자동화와 관련된 컴퓨터과학의 한 분야
- 인공지능 연구자들이 연구하는 문제들과 방법론들의 집합
- 기계를 좀 더 지능적으로 만드는 학문

Turing Test

- 지능이 있다고 가정한 기계의 성능을 사람과 비교하여 평가
- 중요 특징
  - 지능의 객관적인 관념을 제공
  - 혼동되거나 대답할 수 없는 문제의 회피
  - 기계에 비해 살아있는 유기체에 대한 선호를 방지

Turing Test



Chat Bot in KNU

```
nlpdrkim@nlp:~/work/knu_chatbot_sqlite$ ./knu_chatlv1
사 용 자 :
```

Turing Test

튜링테스트의 허와 실

NEXT

"우리는 온라인에서 대화를 나누고 있는 상대방이 실제 인간이라는 전제를 깔고 대화하는 경향이 있습니다."

지난 6월 로봇공학계가 발칵 뒤집히는 사건이 있었다. 러시아의 프로그래머 블라디미르 베셀로프와 우크라이나 출신의 유진 뎀첸코가 공동 개발한 채팅로봇 '유진 구스트만(Eugene Goostman)'이 영국 왕립학회에서 열린 튜링테스트를 통과한 것이다. 튜링테스트는 천재수학자로 불리는 앨런 튜링이 1950년 제안한 인공지능(AI) 컴퓨터 판별 테스트다. 5분간 온라인 채팅을 한 뒤 심사위원의 30%가 인간인지, AI인지를 구분하지 못하면 합격 판정을 받는다.

당시 유진은 자신을 우크라이나에 사는 13세 소년이라 소개하고 채팅을 나눴는데, 33%의 심사위원이 진짜 인간으로 판단해 64년 만에 처음으로 튜링테스트를 통과하는 기염을 토했다. 며칠 뒤 AI 연구자들이 유진은 인간의 대화를 저장해놓은 데이터베이스에서 적당한 문장을 찾아 시뮬레이션 한 것에 불과하다면서 인지능력을 갖춘 AI로 볼 수 없다고 비판하기는 했지만 말이다.

이와 관련 뉴욕대학의 컴퓨터 공학자 어니 데이비스 박사는 이런 평가를 내놓았다.

"사람들은 생각보다 쉽게 속아요. 대화 상대방이 실제 인간이라는 전제를 깔고 대화를 나누는 경향이 있기 때문입니다."

게다가 채팅로봇은 사고능력의 결여를 산만함으로 위장하곤 한다. 예컨대 세계적인 미래학자 레이 커즈와일은 유진에게 이런 질문을 던졌다.

Knowledge needed to build HAL?

- Speech recognition and synthesis
  - Dictionaries (how words are pronounced)
  - Phonetics (how to recognize/produce each sound of English)

- Natural language understanding
  - Knowledge of the English words involved
    - What they mean
    - How they combine (what is a `pod bay door'?)
  - Knowledge of syntactic structure
    - I'm I do, Sorry that afraid Dave I'm can't

What's needed?

- Dialog and pragmatic knowledge
  - "open the door" is a REQUEST (as opposed to a STATEMENT or information-question)
  - It is polite to respond, even if you're planning to kill someone.
  - It is polite to pretend to want to be cooperative (I'm afraid, I can't…)
  - What is 'that' in 'I can't do that'?

- Even a system to book airline flights needs much of this kind of knowledge

Layers of NLP

1. Phonetics & Phonology
2. Morphology
3. Syntax
4. Semantics
5. Pragmatics
6. Discourse & Dialogue

1. Phonetics & Phonology

The study of language sounds, how they are physically formed

dis-k&-'nekt

disconnect

"It is easy to recognize speech."

"It is easy to wreck a nice beach."

2. Morphology

The study of the sub-word units of meaning

disconnect

"not" "to attach"

Even more necessary in some other languages, e.g. Turkish:

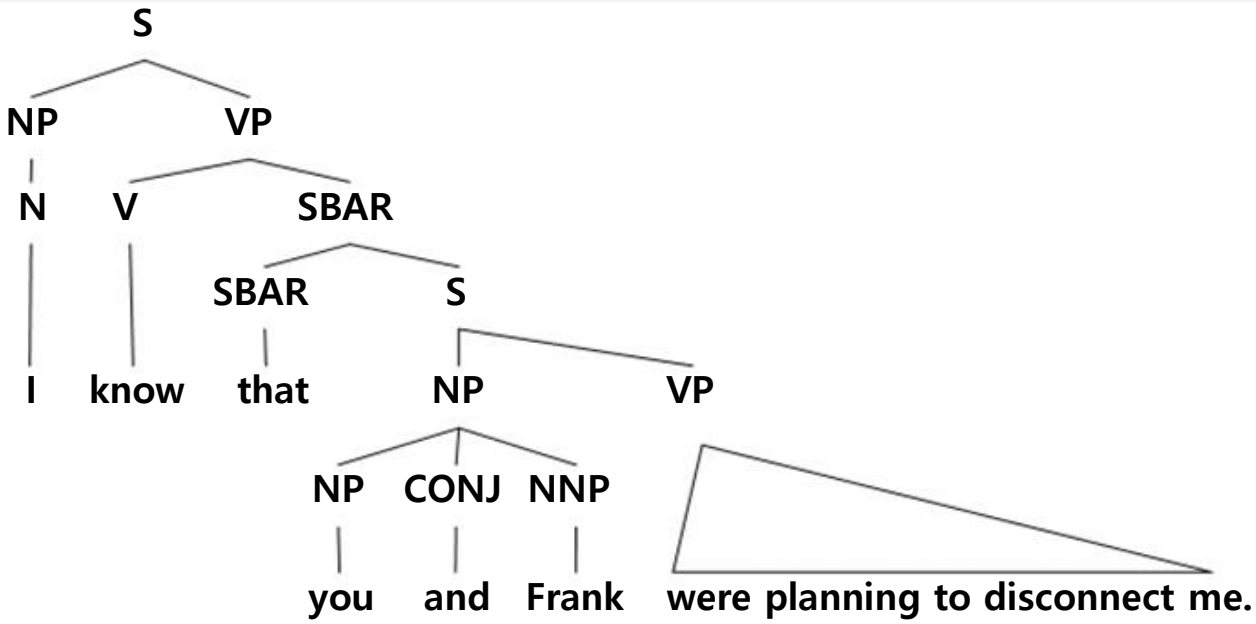
uygarlastiramadiklarimizdanmissinizcasina

uygar las tir ama dik lar imiz dan mis siniz casina

3. Syntax

The study of the structural relationships between words

I know that you and Frank were planning to disconnect me.



Not same structure:

You know me--Frank and I were planning to disconnect that.

4. Semantics

The study of the literal meaning

I know that you and Frank were planning to disconnect me.

ACTION = disconnect
ACTOR = you and Frank
OBJECT= me

5. Pragmatics

The study of how language is used to accomplish goals

What should you conclude from the fact I said something?
How should you react?

I'm sorry Dave, I'm afraid I can't do that.

→ Includes notions of polite and indirect styles

6. Discourse & Dialogue

The study of linguistic units larger than a single utterance

The structure of conversations: turn taking, thread of meaning

David Bowman:
Open the pod bay doors, Hal.
HAL:
I'm sorry, Dave, I'm afraid I can't do that.
David Bowman:
What are you talking about, Hal?
...HAL:
I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

NLP Demo in KNU

What can NLP do? Question-answering

Watson and the Jeopardy!

Ambiguity

- Computational linguists are obsessed with ambiguity
- Ambiguity is a fundamental problem of computational linguistics
- Resolving ambiguity is a crucial goal
- Find at least 5 meanings of this sentence:
  - I made her duck

Ambiguity

- Find at least 5 meanings of this sentence: I made her duck
  - I cooked waterfowl for her benefit (to eat)
  - I cooked waterfowl belonging to her
  - I created the (plaster?) duck she owns
  - I caused her to quickly lower her head or body
  - I waved my magic wand and turned her into undifferentiated waterfowl
- At least one other meaning that's inappropriate for gentle company

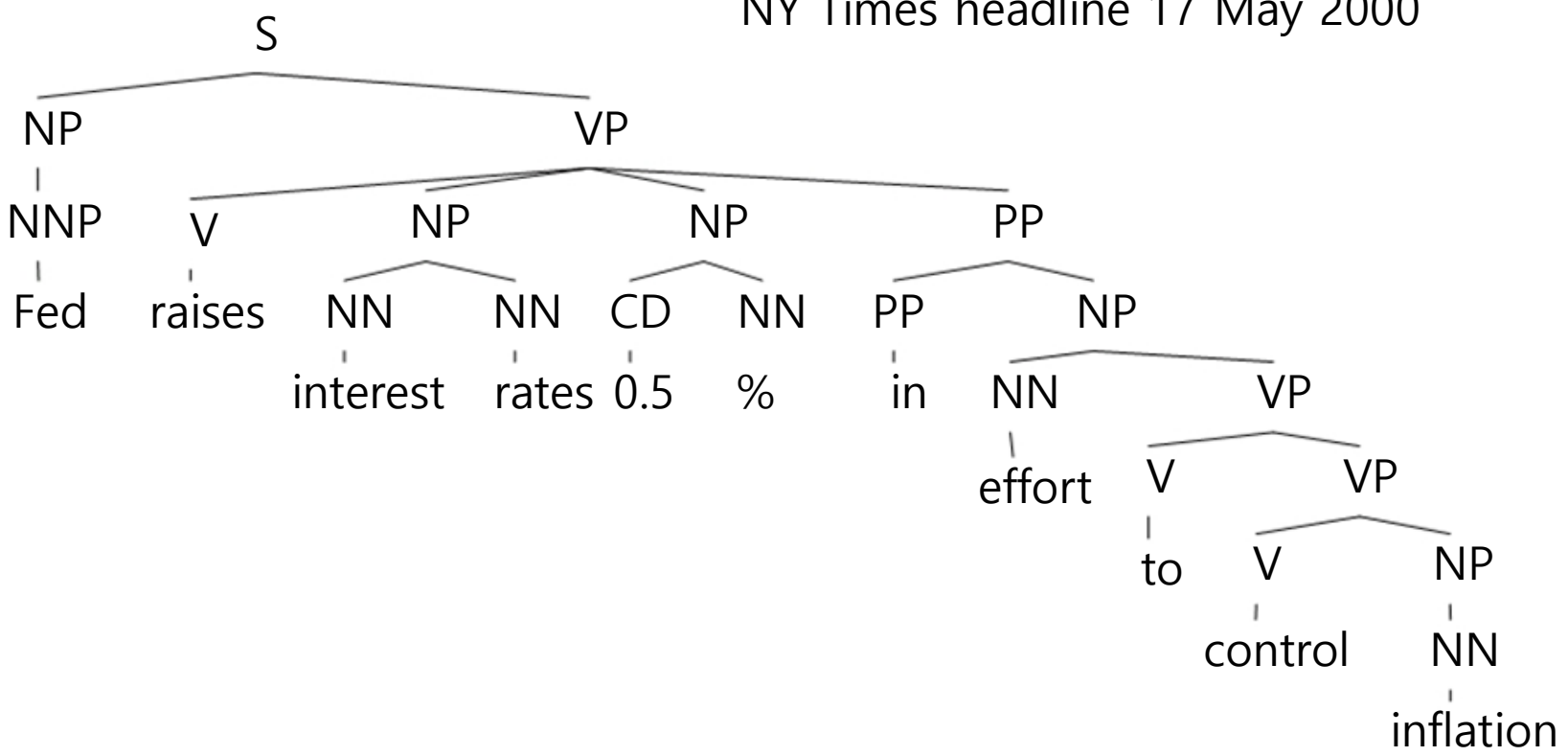
Ambiguity is Pervasive

- Ambiguity in Phonetics
  - I mate or duck
  - I'm eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I'm aid her duck
  - I mate her duck
  - I'm ate her duck
  - I'm ate or duck
  - I mate or duck

Ambiguity is Pervasive

Fed raises interest rates 0.5% in effort to control inflation

NY Times headline 17 May 2000



Ambiguity is Pervasive

Part-of-speech ambiguities

Syntactic attachment ambiguities

	VBZ	VB VBZ	VBZ			
NNP	NNS	NNS	NNS	CD	NN	
Fed	raises	interest	rates	0.5	%	in effort to control inflation

Word sense ambiguities:

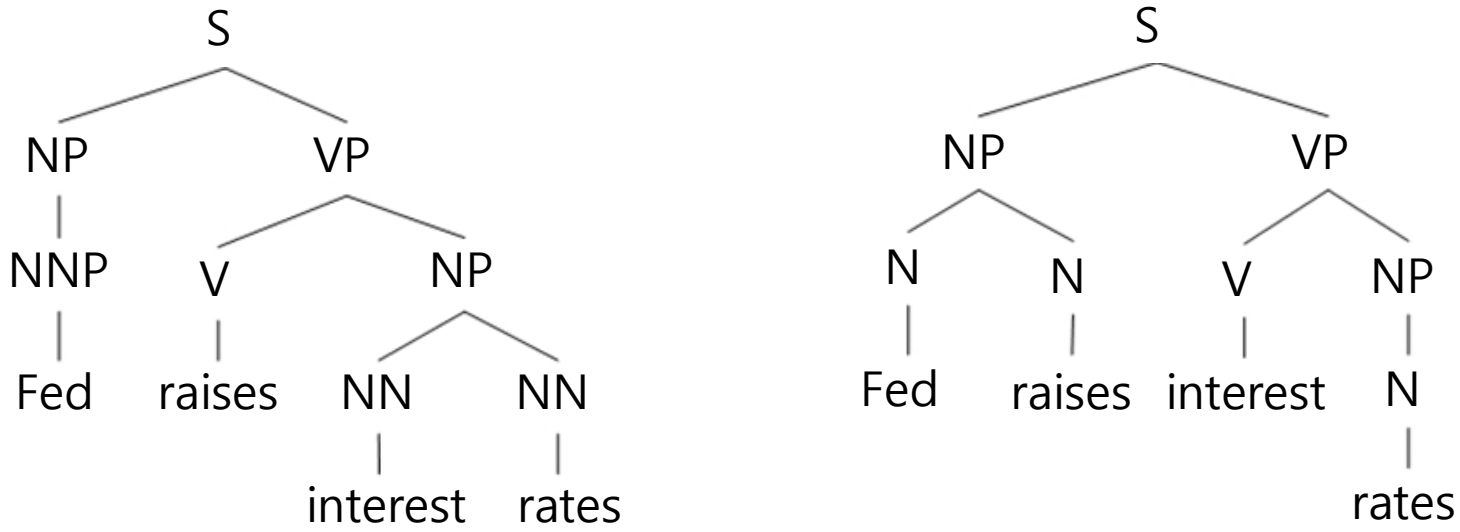
Fed → "federal agent"

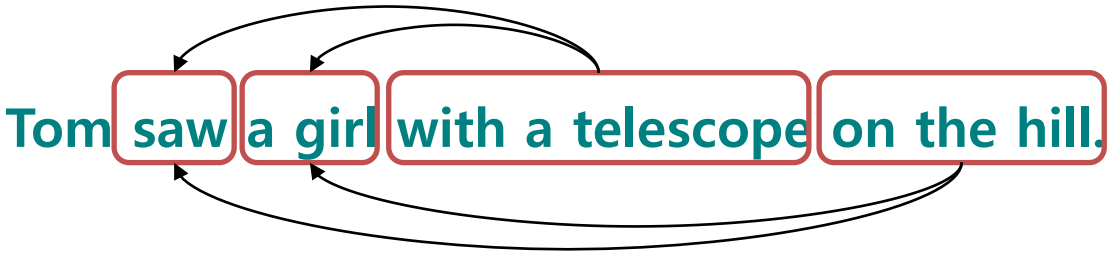
interest → a feeling of wanting to know or learn more

Semantic interpretation ambiguities above the word level.

Ambiguity is Pervasive

Syntactic attachment ambiguity





Probabilistic Models of Language

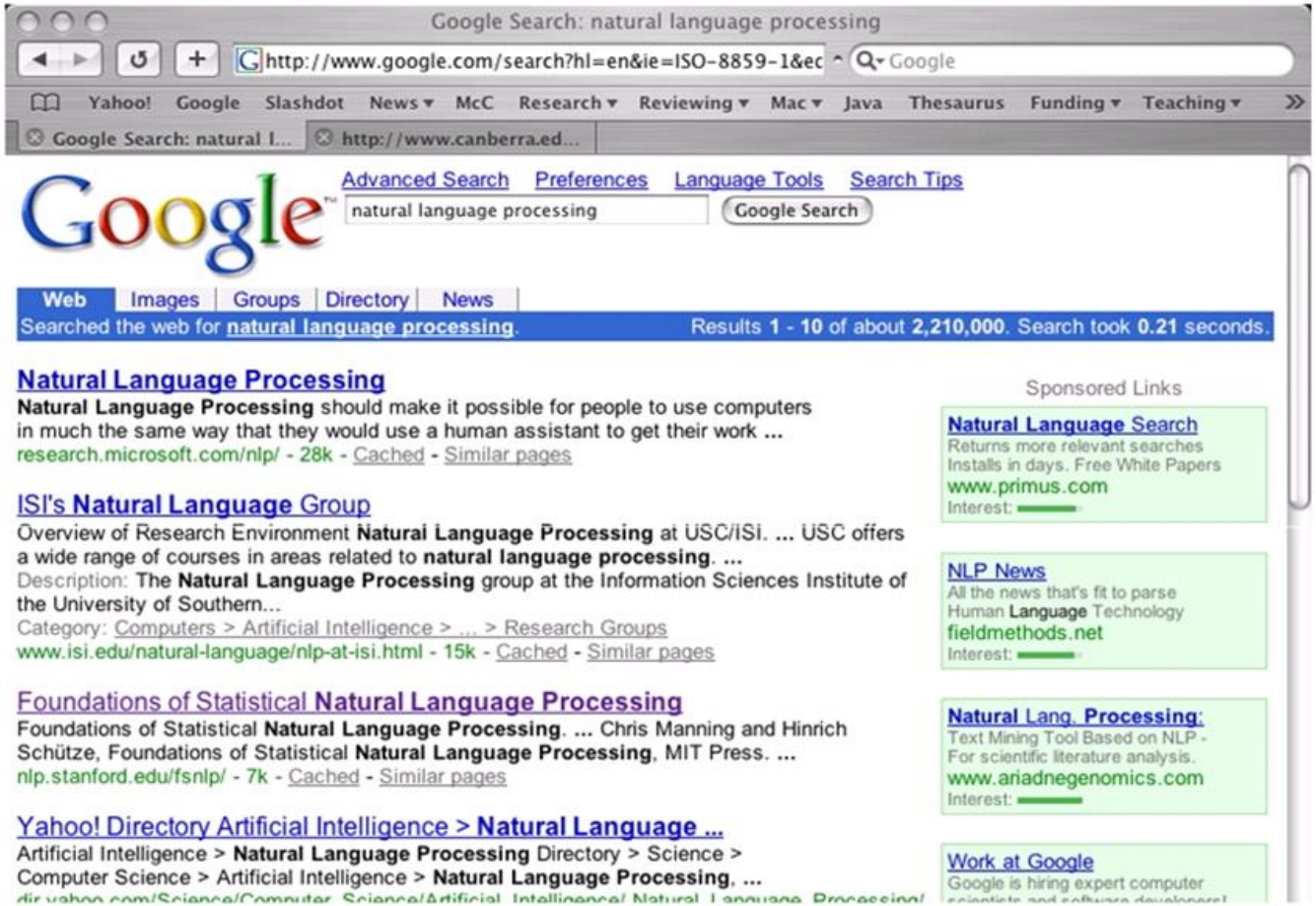
- To handle this ambiguity and to integrate evidence from multiple levels
- The tools of probability:
  - Bayesian Classifiers (not rules)
  - Hidden Markov Models (not DFAs)
  - Probabilistic Context Free Grammars
  - Other tools of Machine Learning, AI, Statistics

Where is NLP?

- Goals can be very far-reaching
  - True text understanding
  - Reasoning and decision-making from text
  - Real-time spoken dialogue
- Goals can be very down-to-earth
  - Searching the Web
  - Context-sensitive spelling correction
  - Analyzing reading-level or authorship statistically
  - Extracting company names and locations from news articles

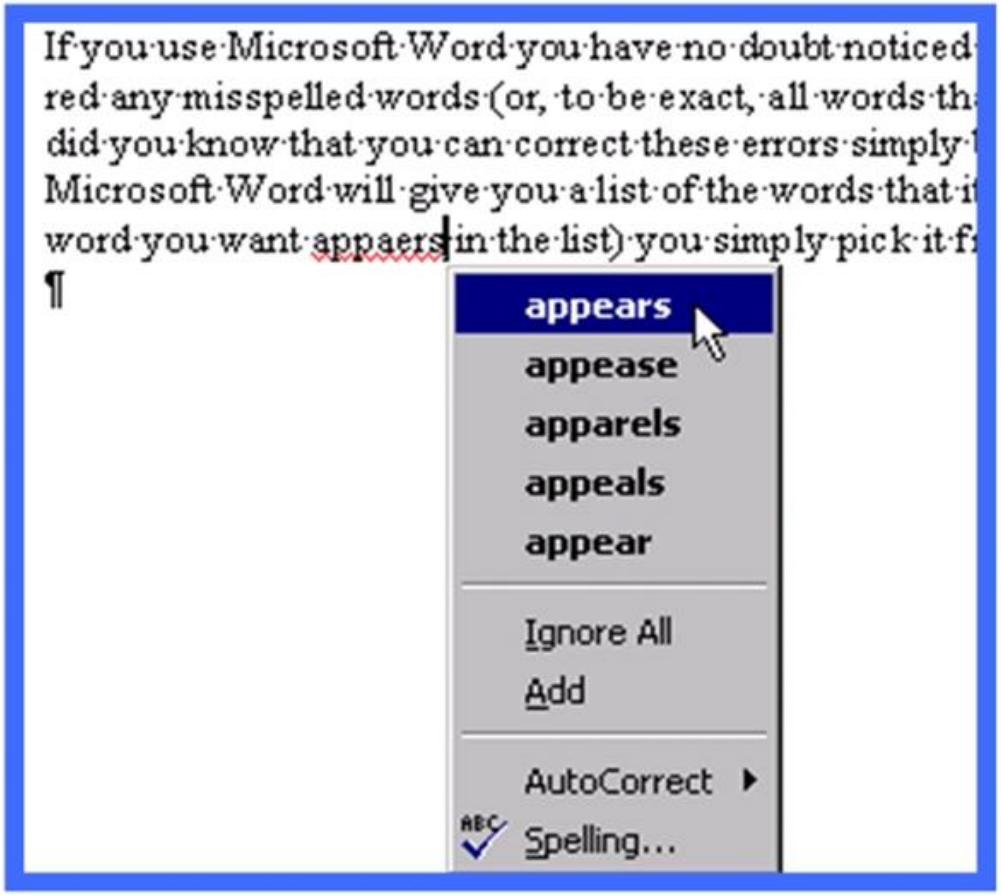
Example Applications of NLP

Information retrieval



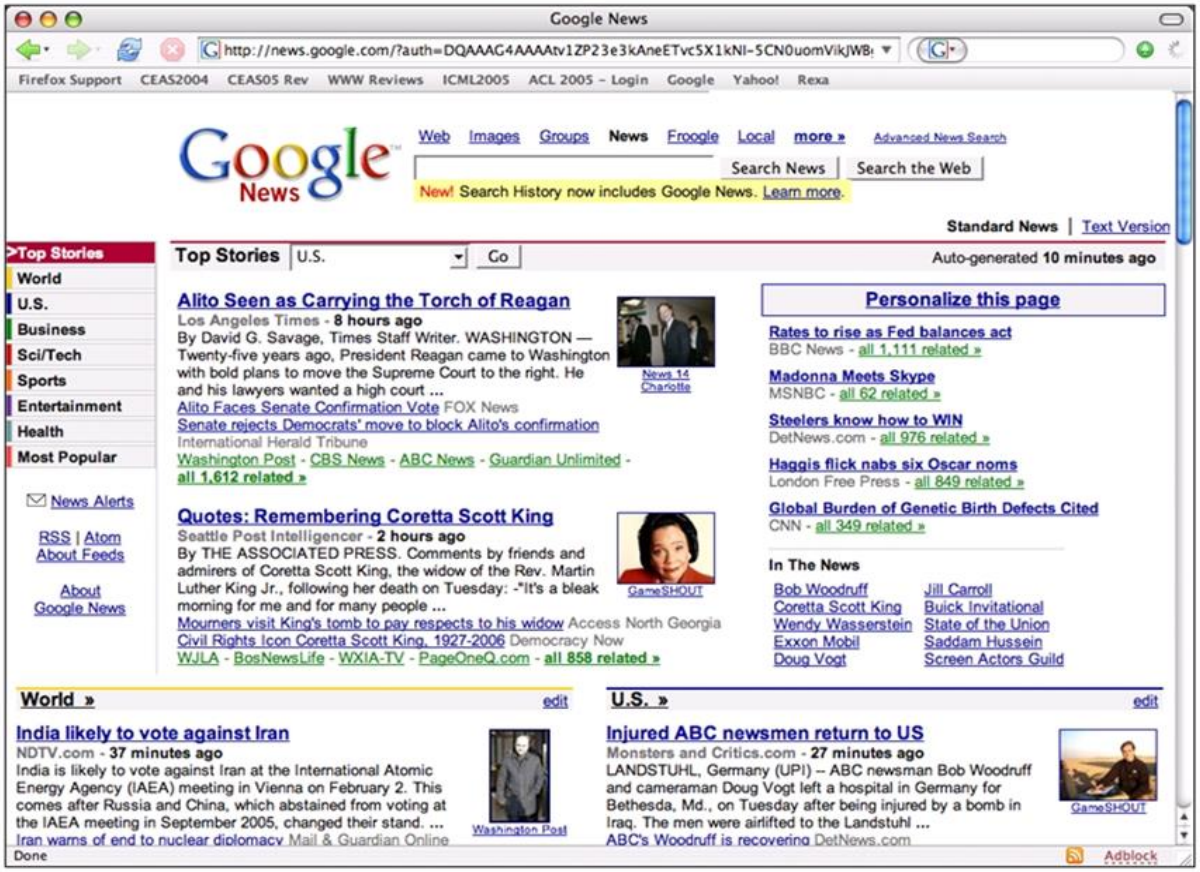
Example Applications of NLP

MSWord spelling correction, grammar checking



Example Applications of NLP

News categorization and summarization



Example Applications of NLP

Information Extraction: Find experts, employees

http://www.zoominfo.com

Dr. Andrew McCallum
Action Editor
Journal of Machine Learning Research
Last Mentioned on 10/12/2003

Actions
Send This Profile
Update Your Profile
Email Not Available

Other Titles Held:
Member, Editorial Board

Additional Current Employment	
Carnegie Mellon University	Post-Doctoral Fellow
	Adjunct Faculty Member
	Adjunct Faculty Position
University of Massachusetts Amherst, CO	Research Associate Professor
Adjunct Faculty	Research Scientist

Board Memberships and Affiliations	
Intelliseek Inc	Member of Advisory Board
IJCAI	Member, Program Committees (past)
AAAI	Member, Program Committees (past)
ICML	Member, Program Committees (past)
NIPS	Member, Program Committees (past)

Past Employment History	
WhizBang Labs Inc	Vice President of Research and Development
Just Research	Research Scientist
Biomedical Information Communication Center of Oregon Health Sciences University	Machine Learning Researcher

Education		
University of Rochester	Ph.D.	Computer Science
Dartmouth College	Bachelor of Arts	Computer Science

Information about Andrew McCallum was compiled from 6 sources:

JMLR Inc
http://www.jmlr.org
JMLR, which publishes high-quality scholarly articles in all areas of machine learning, competes with the commercial journal Machine Learning, which costs US$1006. A number of Machine Learning editorial board members have resigned to join the editorial board of JMLR. ... (more)
Click here to find other people who work for *JMLR Inc*

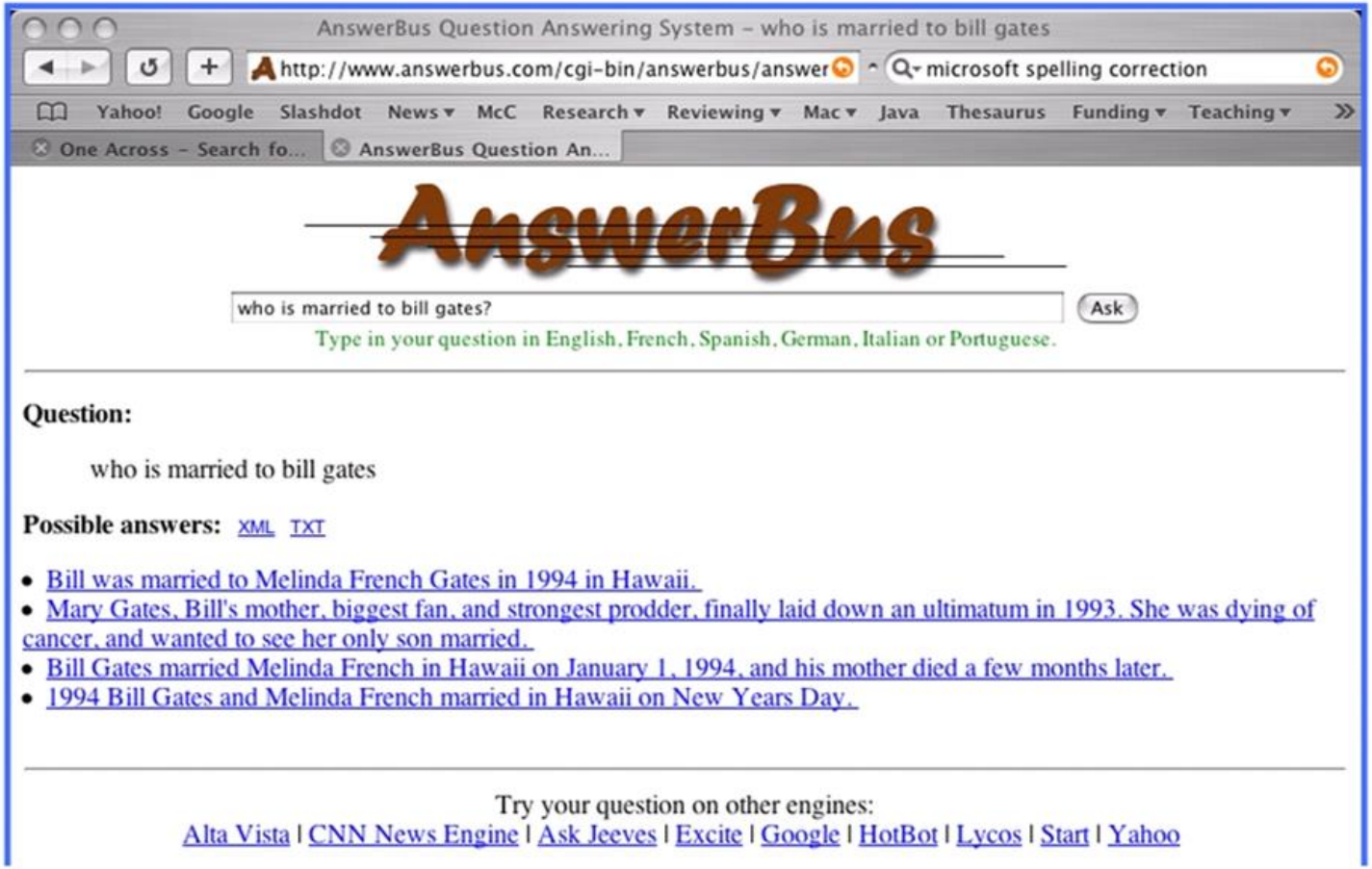
WhizBang Labs Inc
Contact Us Corporate Headquarters
3210 North Canyon Road Suite 200
Provo, UT 84604
Phone: (801) 418-7100
Fax: (801) 818-0300
http://www.whizbanglabs.com
WhizBang! Labs, founded in 1999, is a leader in the field of information extraction and document auto-tagging from unstructured data sources.
Through our products and services, we analyze unstructured content in both on-line and off-line formats, locate and extract key data elements into XML-tagged ... (more)
Click here to find other people who work for *WhizBang Labs Inc*

Intelliseek Inc
1128 Main Street , Fourth Floor
Cincinnati, OH 45202-7236
Phone: 513-618-6700

Go to "http://networking2.eliyon.com/Networking/default.asp"

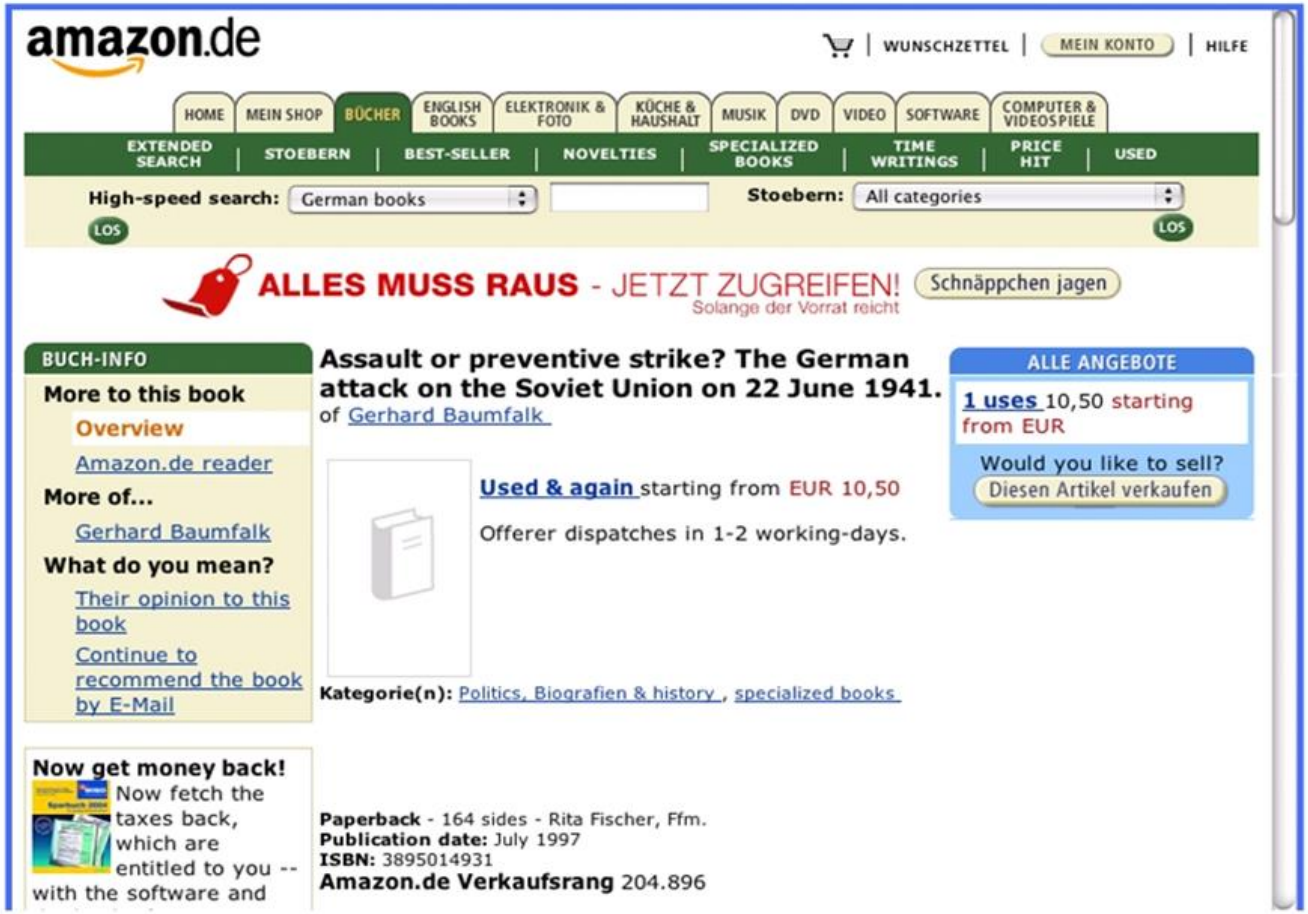
Example Applications of NLP

Question-answering



Example Applications of NLP

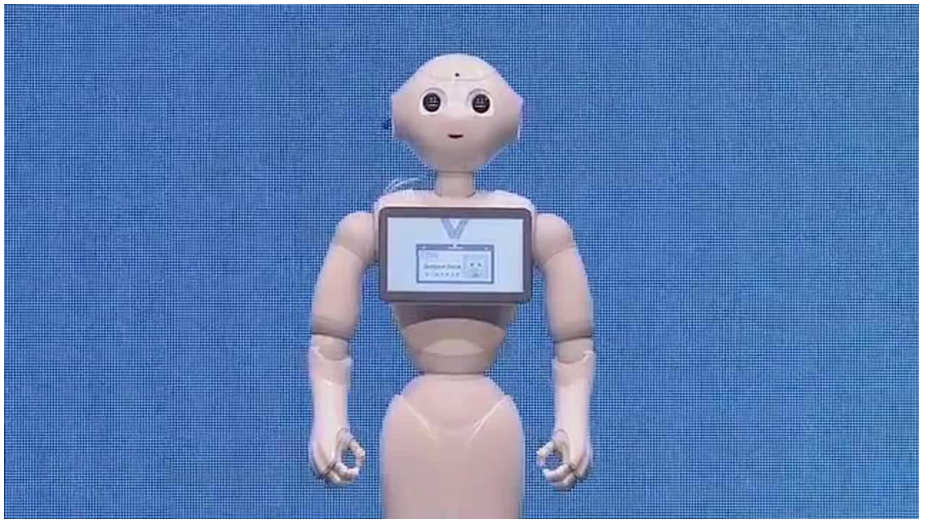
Machine translation



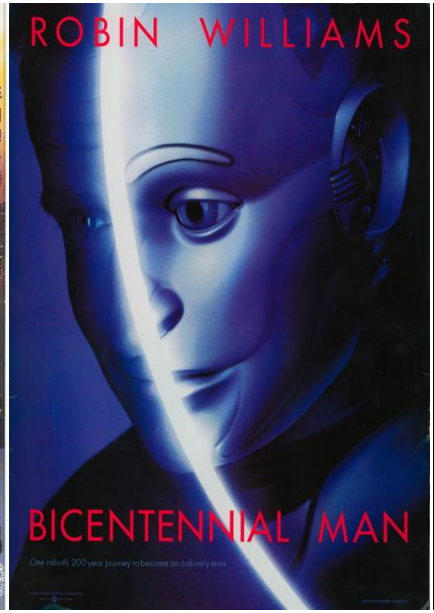
What can NLP do? Robot

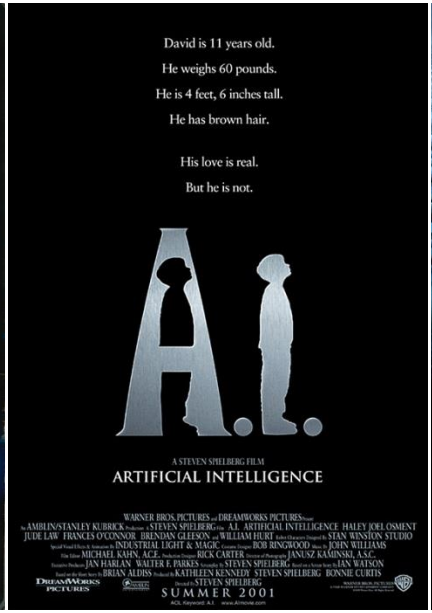
Amazon echo

Pepper with IBM Watson



NEXT? ChatGPT?





확인 문제

- 다음 각 문장을 이해함에 있어서 주어진 분석 단계에서 어떤 사실을 밝혀 낼 수 있는지 설명하시오.
  - [형태소 분석] 나는 과자를 먹었다.
  - [구문 분석] 나를 과자를 먹었다.
  - [의미 분석] 나는 자동차를 먹었다.
  - [담화 분석] 나는 자동차를 먹었다. 그건 너무 맛있어.

KU KONKUK UNIVERSITY

KAIST

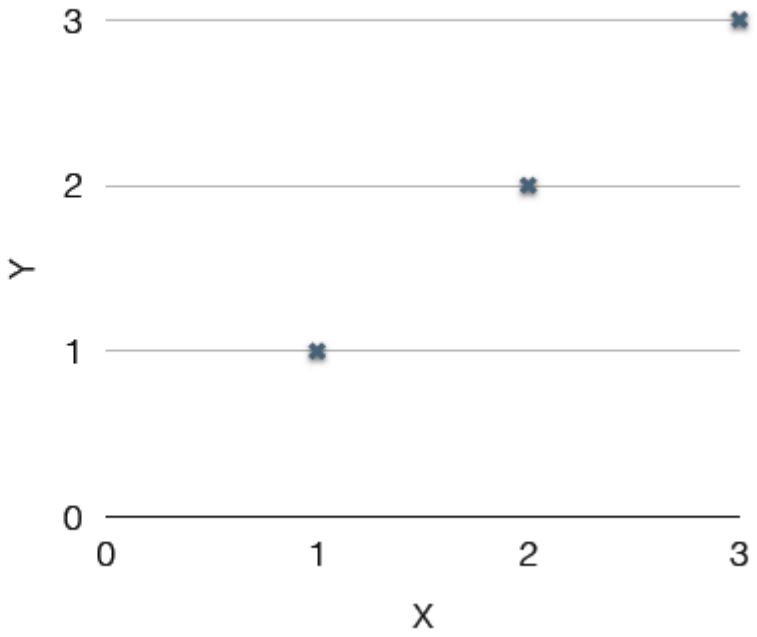
Concept of Machine Learning

자연어처리연구실

본 챕터는 김성훈 박사님의 "모두를 위한 머신러닝/딥러닝 강의 (https://hunkim.github.io/ml/)" 자료를 참고하여 편집되었습니다.

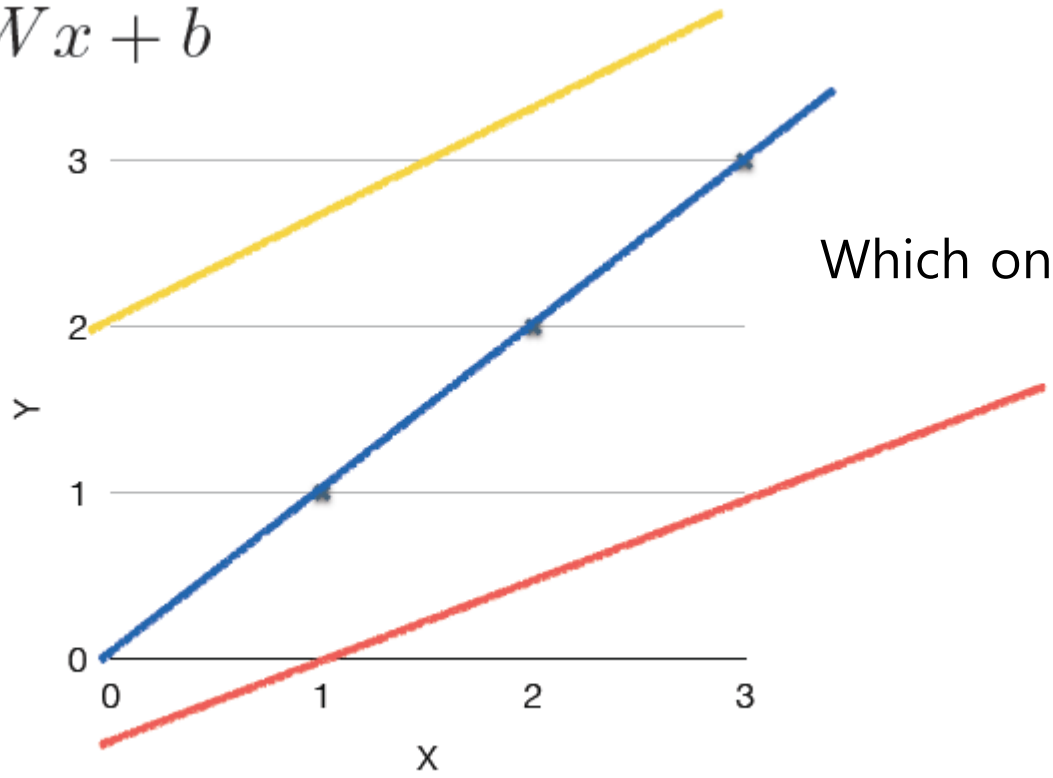
What is Learning?

x	Y
1	1
2	2
3	3



Linear Hypothesis

$$H(x) = Wx + b$$



Which one is better?

Multi-variable

$$H(x_1, x_2) = w_1 x_1 + w_2 x_2 + b$$

$$H(x_1, x_2, x_3, ..., x_n) = w_1 x_1 + w_2 x_2 + w_3 x_3 + ... + w_n x_n + b$$

Matrix representation

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + ... + w_n x_n$$

$$\begin{bmatrix} w1 & w2 & w3 \end{bmatrix} \times \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} w1 \times x1 + w2 \times x2 + w3 \times x3 \end{bmatrix}$$

Matrix Representation

$$\begin{bmatrix} w1 & w2 & w3 \end{bmatrix} \times \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} w1 \times x1 + w2 \times x2 + w3 \times x3 \end{bmatrix}$$

$$H(X) = WX + b$$

With b vector

$$\begin{bmatrix} b & w1 & w2 & w3 \end{bmatrix} \times \begin{bmatrix} 1 \\ x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} b \times 1 + w1 \times x1 + w2 \times x2 + w3 \times x3 \end{bmatrix}$$

$$H(X) = WX$$

Without b vector

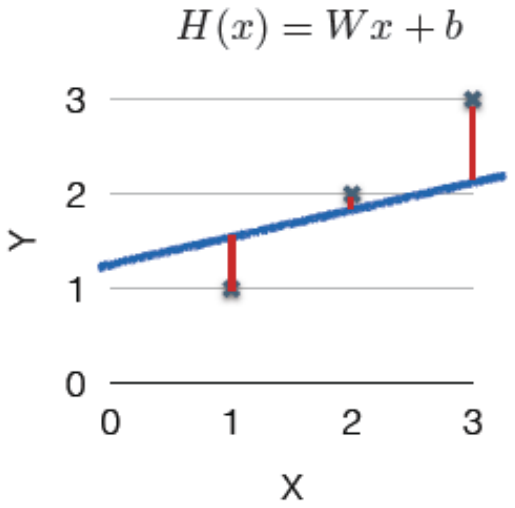
$$H(X) = W^T X$$

Transpose representation

Which hypothesis is better?

- How fit the line to our (training) data

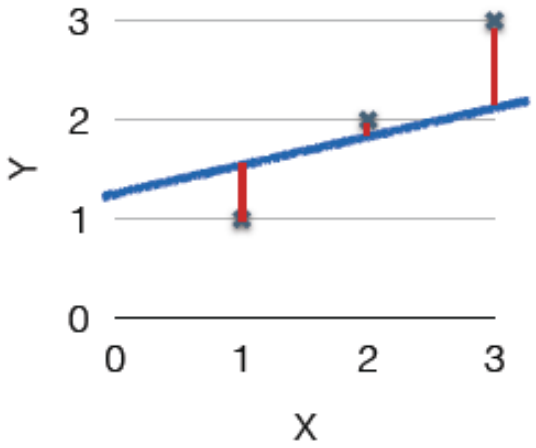
$$H(x) - y$$



Cost Function

$$\frac{(H(x^{(1)}) - y^{(1)})^2 + (H(x^{(2)}) - y^{(2)})^2 + (H(x^{(3)}) - y^{(3)})^2}{3}$$

$$H(x) = Wx + b$$



$$cost = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

Our goal? $\underset{W,b}{\text{minimize}}\, cost(W, b)$

Cost function을 최소로 하는hypothesis가 무엇일까?

Hypothesis and Cost

$$H(x) = Wx + b$$

$$cost(W, b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

Simplifying without b vector

$$H(x) = Wx$$

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

What cost(W) looks like?

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

x	Y
1	1
2	2
3	3

- W=1, cost(W)=0

$$\frac{1}{3}((1*1-1)^2 + (1*2-2)^2 + (1*3-3)^2)$$

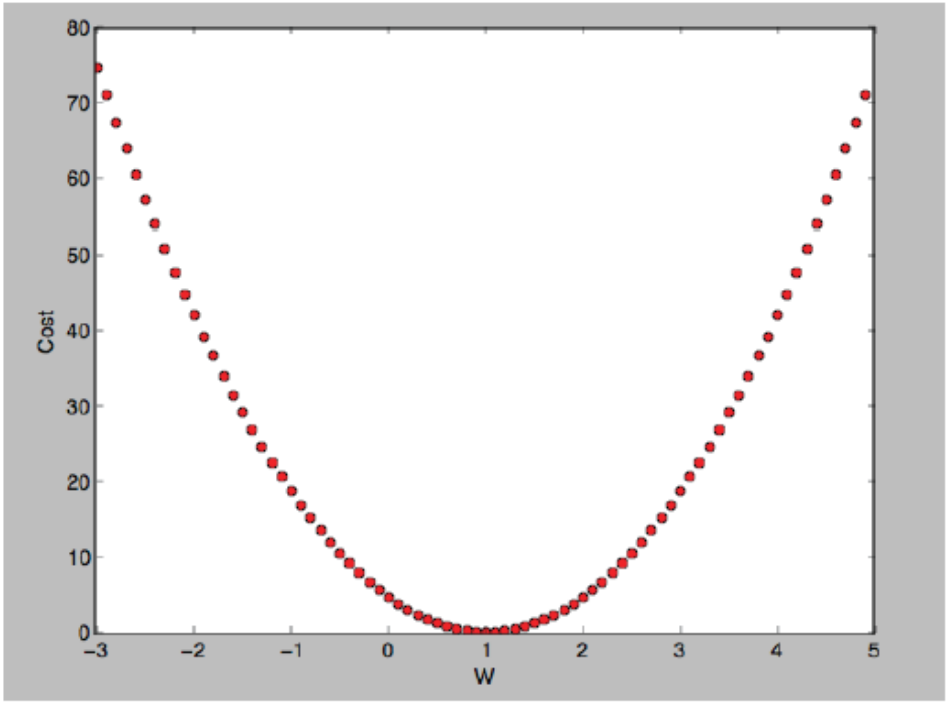
- W=0, cost(W)=4.67

$$\frac{1}{3}((0*1-1)^2 + (0*2-2)^2 + (0*3-3)^2)$$

- W=2, cost(W)=? ?

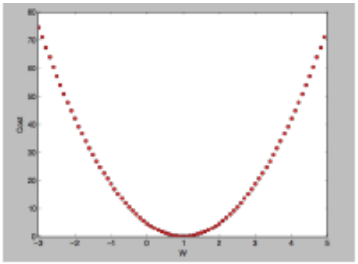
What cost(W) looks like?

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$



How to Minimize Cost?

= How to find the lowest point?



- Start with initial guesses
  - Start at 0,0 (or any other value)
  - Keeping changing W and b a little bit to try and reduce cost(W, b)
- Each time you change the parameters, you select the gradient which reduces cost(W, b) the most possible
- Repeat
- Do so until you converge to a local minimum
- Has an interesting property
  - Where you start can determine which minimum you end up

Formal Definition of Gradient Decent

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

$$cost(W) = \frac{1}{2m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

$$W := W - \alpha\frac{\partial}{\partial W}\frac{1}{2m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

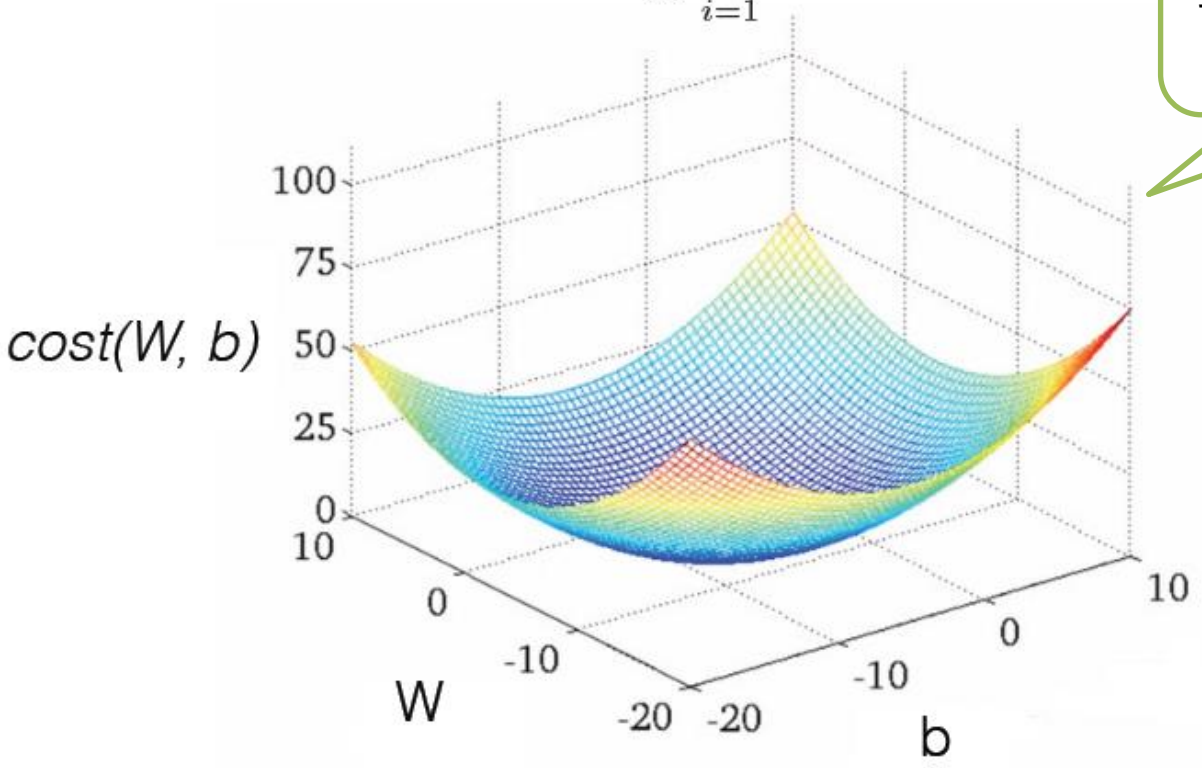
$$W := W - \alpha\frac{\partial}{\partial W}cost(W)$$

$$W := W - \alpha\frac{1}{2m}\sum_{i=1}^{m}2(Wx^{(i)} - y^{(i)})x^{(i)}$$

$$W := W - \alpha\frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})x^{(i)}$$

Convex Function

$$cost(W,b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$



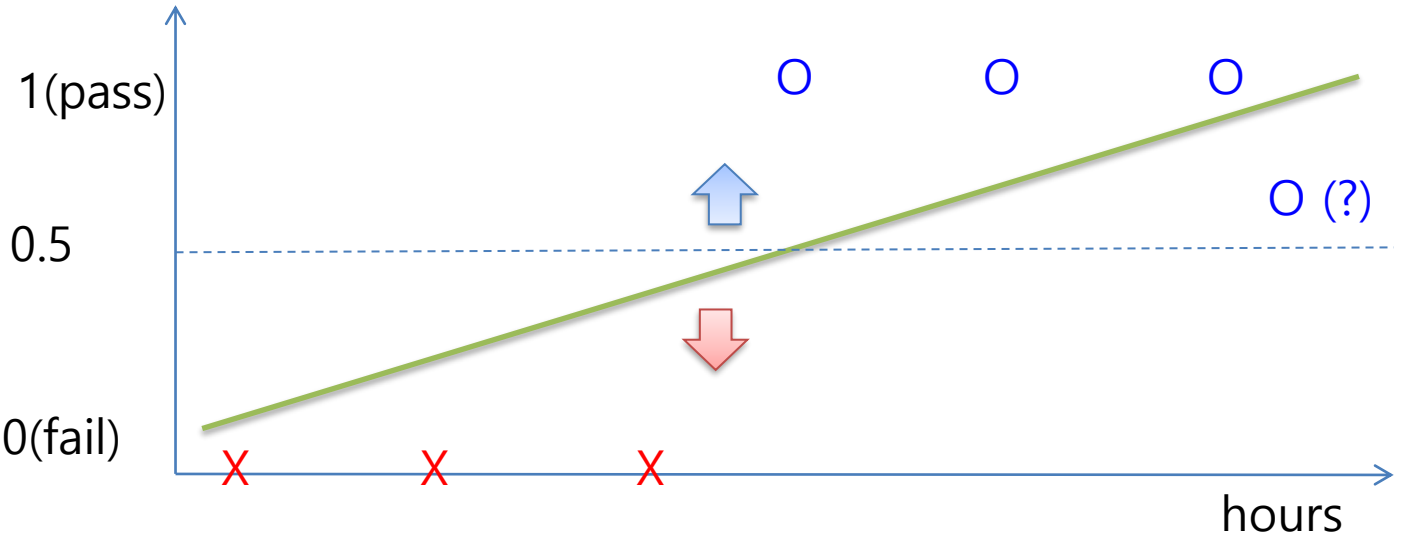
Learning → Cost를 최소화하는 W와 b를 찾자!

Regression to Classification

Classification problems

- Spam Detection: Spam (1) or Ham (0)
- Facebook feed: show(1) or hide(0)
- Credit Card Fraudulent Transaction detection: legitimate(0) or fraud (1)

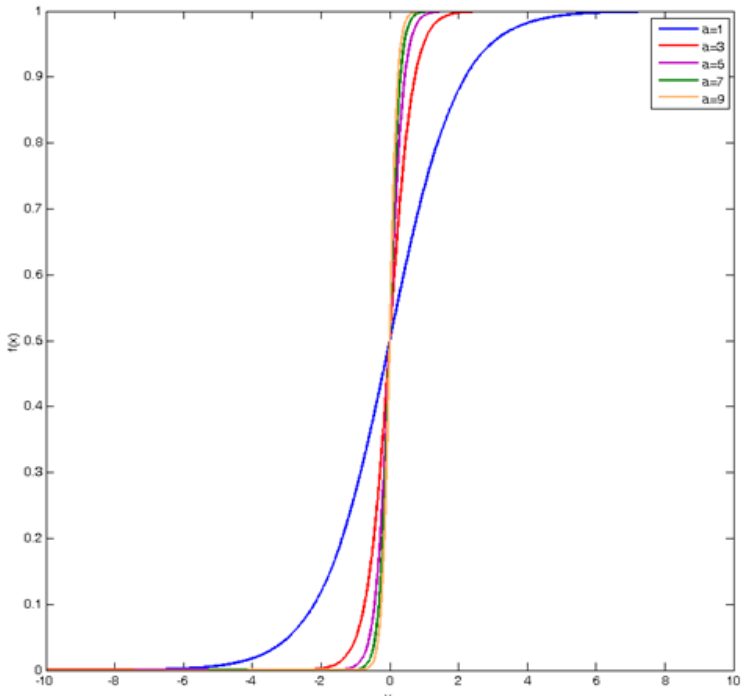
Pass/Fail based on study hours?



Logistic Hypothesis

$$H(x) = Wx + b \quad \Rightarrow \quad g(z) = \frac{1}{(1 + e^{-z})}$$

WHY?
0과 1 사이 값으로 변환



Logistic Hypothesis & Cost Function

$$H(X) = \frac{1}{1 + e^{-W^T X}}$$

$$cost(W, b) = \frac{1}{m} \sum_{i=1}^{m} (H(x^{(i)}) - y^{(i)})^2$$

Many local minimums

New Cost Function

$$Cost(W) = \frac{1}{m} \sum c(H(x), y)$$

$$c(H(x), y) = \begin{cases} -log(H(x)) & : y = 1 \\ -log(1 - H(x)) & : y = 0 \end{cases}$$

H(x)=1일 때 C값은? ?

Cost Function

$$Cost(W) = \frac{1}{m}\sum c(H(x), y)$$

$$c(H(x), y) = \begin{cases} -log\big(H(x)\big) & : y = 1 \\ -log\big(1 - H(x)\big) & : y = 0 \end{cases}$$

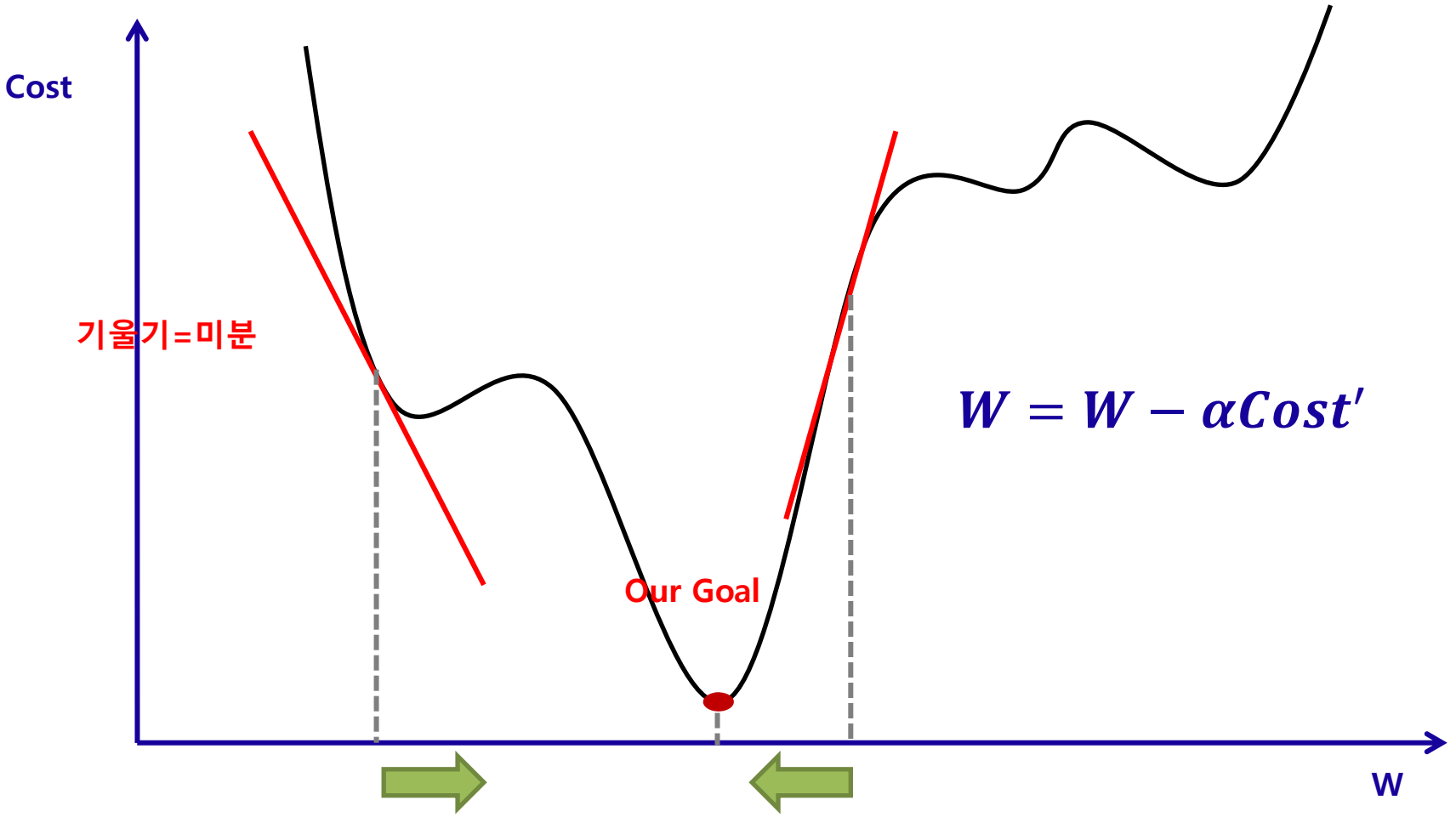
$$c(H(x), y) = -ylog\big(H(x)\big) - (1 - y)log(1 - H(x))$$

Minimize Cost → Gradient decent algorithm

$$Cost(w) = -\frac{1}{m}\sum ylog\big(H(x)\big) + (1 - y)log(1 - H(x))$$

$$W := W - \alpha\frac{\partial}{\partial W}cost(W)$$

Goal of ML Models



확인 문제

- 다음 학습 데이터와 비용 함수(cost function)가 주어지고, 초기 W값이 2이고 학습률이 0.1일 때, gradient decent 알고리즘에 의해 1회 학습 후 수정된 W 값을 구하시오.

[학습 데이터]

X (입력)	Y (출력)
1	1
2	3
3	5

[비용 함수]

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

확인 문제

?

KU KONKUK UNIVERSITY

KAIST

Artificial Neural Network

자연어처리연구실

본 챕터의 일부는 김성훈 박사님의 “모두를 위한 머신러닝/딥러닝 강의 (https://hunkim.github.io/ml/)” 자료를 참고하여 편집하였습니다.

ANN (Artificial Neural Networks)

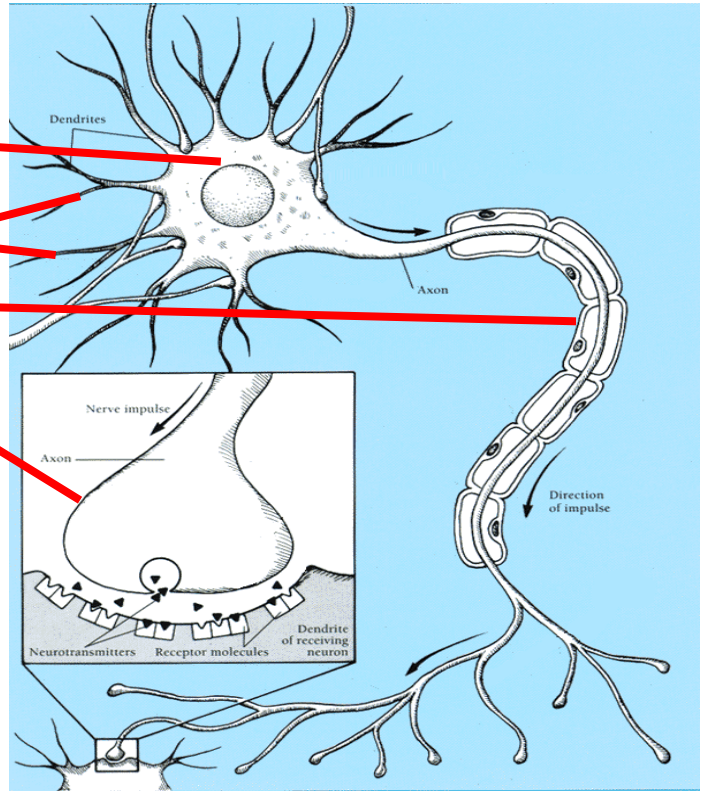
- 수학적 논리학이 아닌 인간의 두뇌를 모방하여 수많은 간단한 처리기들(뉴런)의 네트워크를 통해 문제를 해결하는 기계학습 모델

Cell body

Dendrites(수상돌기)

Axon(축색돌기)

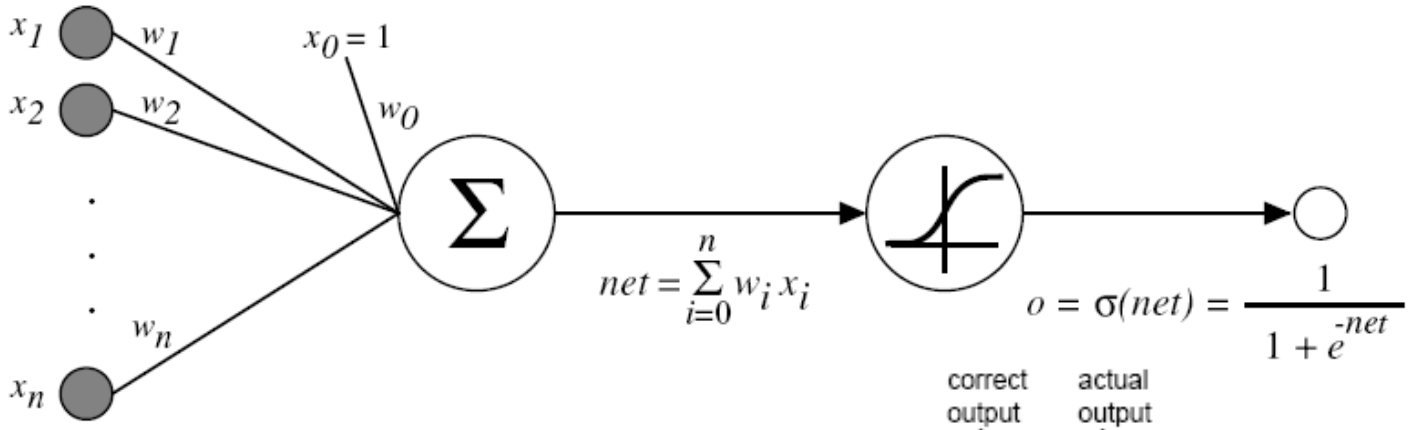
Synaptic terminals(시냅스)



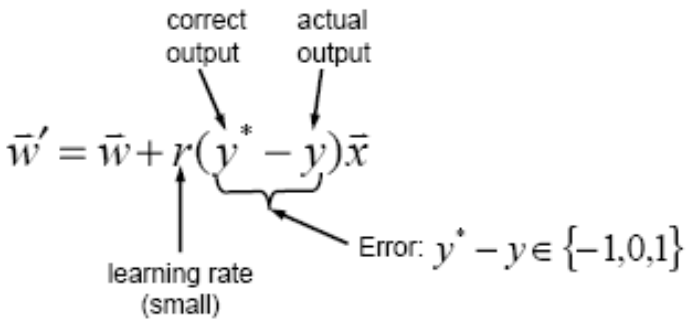
ANN (Artificial Neural Networks)

- 수학적 논리학이 아닌 인간의 두뇌를 모방하여 수많은 간단한 처리기들(뉴런)의 네트워크를 통해 문제를 해결하는 기계학습 모델

Dendrites(수상돌기) Axon(축색돌기) Synaptic terminals(시냅스)



학습(델타룰): 정답과 출력을 비교하여 그 차이를 가중치 조정에 반영



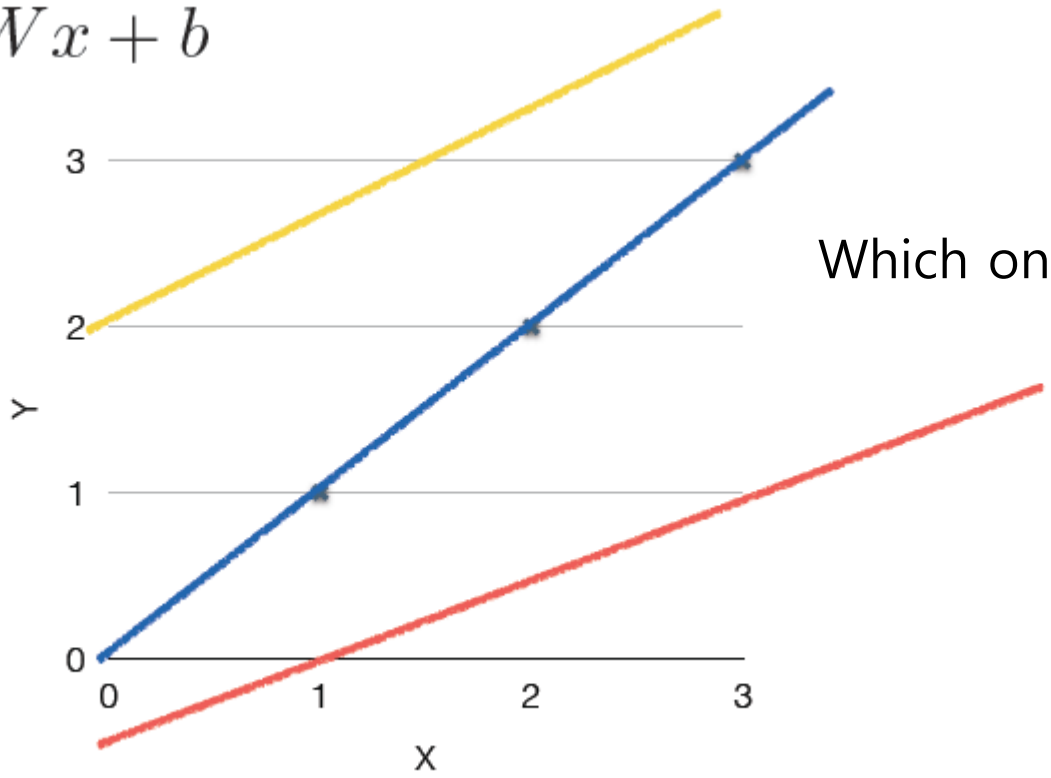
Brief ANN History

- Frank Rosenblatt, 1957
  - Single-layer perceptron
- Minsky & Papert 1969
  - ANN is a linear function (1st winter season)
- Rumelhart, Hinton & Williams, 1986
  - Back propagation algorithm for Multi-layer perceptron
  - Vanishing gradient problem! (2nd winter season)
- Geoffrey Hinton, 2009 → Yoshua Bengio, Andrew Ng, Ian Goodfellow
  - New activation function, ReLU, for deep neural networks
  - Drop-out for increasing robustness

REMIND

Linear Hypothesis

$$H(x) = Wx + b$$



Which one is better?

REMIND

Matrix Representation

$$\begin{bmatrix} w1 & w2 & w3 \end{bmatrix} \times \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} w1 \times x1 + w2 \times x2 + w3 \times x3 \end{bmatrix}$$

$$H(X) = WX + b$$

With b vector

$$\begin{bmatrix} b & w1 & w2 & w3 \end{bmatrix} \times \begin{bmatrix} 1 \\ x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} b \times 1 + w1 \times x1 + w2 \times x2 + w3 \times x3 \end{bmatrix}$$

$$H(X) = WX$$

Without b vector

$$H(X) = W^T X$$

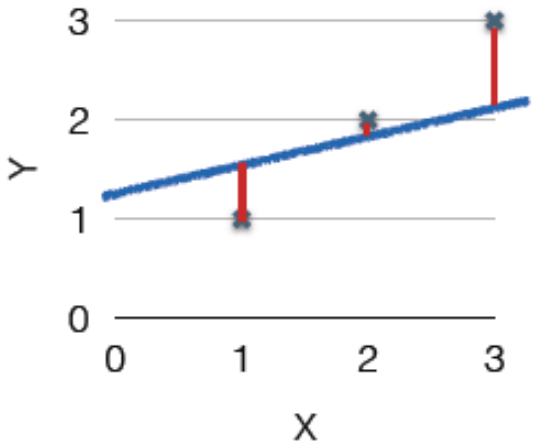
Transpose representation

REMIND

Cost Function

$$\frac{(H(x^{(1)}) - y^{(1)})^2 + (H(x^{(2)}) - y^{(2)})^2 + (H(x^{(3)}) - y^{(3)})^2}{3}$$

$$H(x) = Wx + b$$



$$cost = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

Our goal? $\underset{W,b}{\text{minimize}}\, cost(W, b)$

Cost function을 최소로 하는hypothesis가 무엇일까?

REMIND

Formal Definition of Gradient Decent

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

$$cost(W) = \frac{1}{2m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

$$W := W - \alpha\frac{\partial}{\partial W}\frac{1}{2m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

$$W := W - \alpha\frac{\partial}{\partial W}cost(W)$$

$$W := W - \alpha\frac{1}{2m}\sum_{i=1}^{m}2(Wx^{(i)} - y^{(i)})x^{(i)}$$

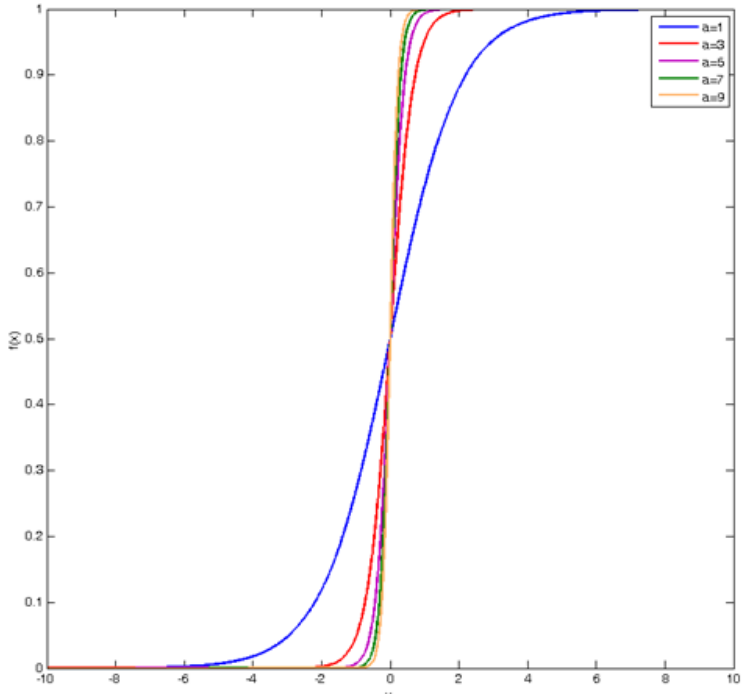
$$W := W - \alpha\frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})x^{(i)}$$

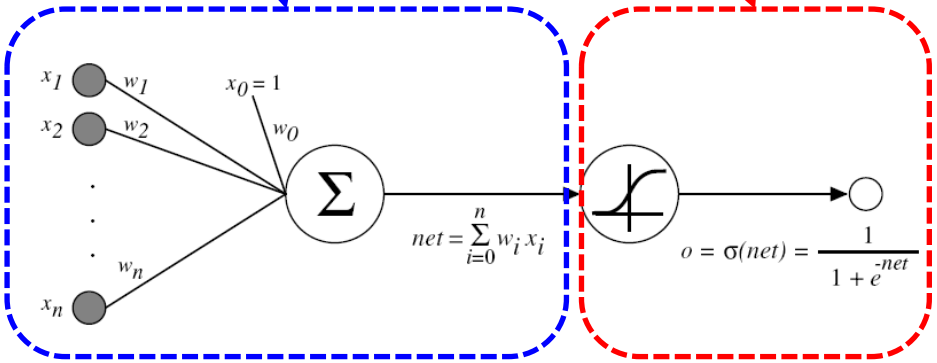
REMIND

Logistic Hypothesis

$$H(x) = Wx + b \quad \Rightarrow \quad g(z) = \frac{1}{(1 + e^{-z})}$$

WHY?
0과 1 사이 값으로 변환





Architecture of ANN

REMIND

Cost Function

$$Cost(W) = \frac{1}{m}\sum c(H(x), y)$$

$$c(H(x), y) = \begin{cases} -log\big(H(x)\big) & : y = 1 \\ -log\big(1 - H(x)\big) & : y = 0 \end{cases}$$

$$c(H(x), y) = -ylog\big(H(x)\big) - (1 - y)log(1 - H(x))$$

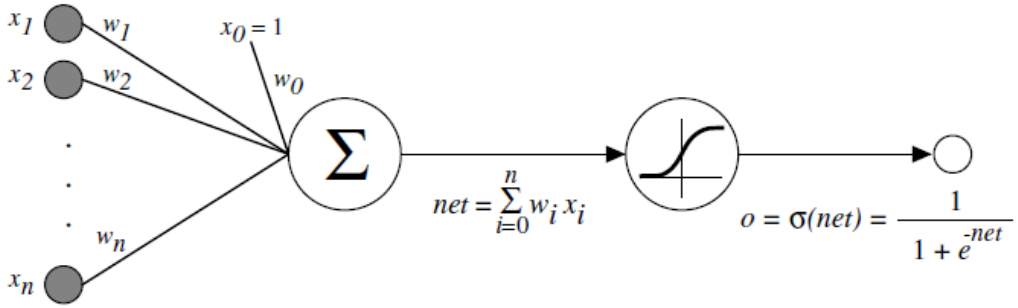
Minimize Cost → Gradient decent algorithm

$$Cost(w) = -\frac{1}{m}\sum ylog\big(H(x)\big) + (1 - y)log(1 - H(x))$$

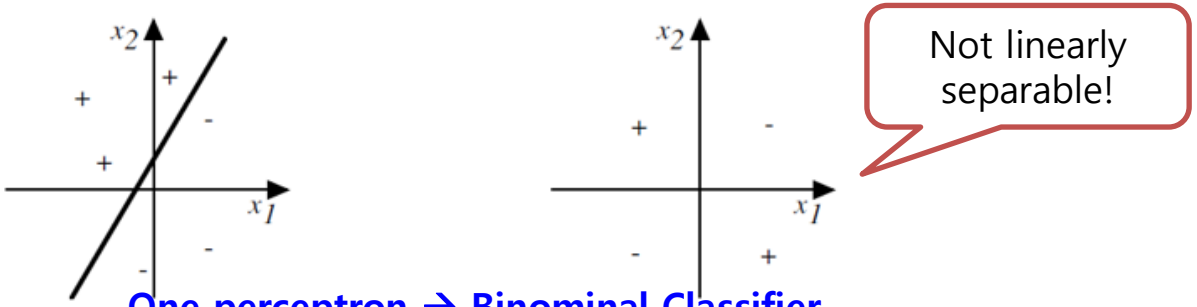
$$W := W - \alpha\frac{\partial}{\partial W}cost(W)$$

퍼셉트론 (Perceptron)

- 구조



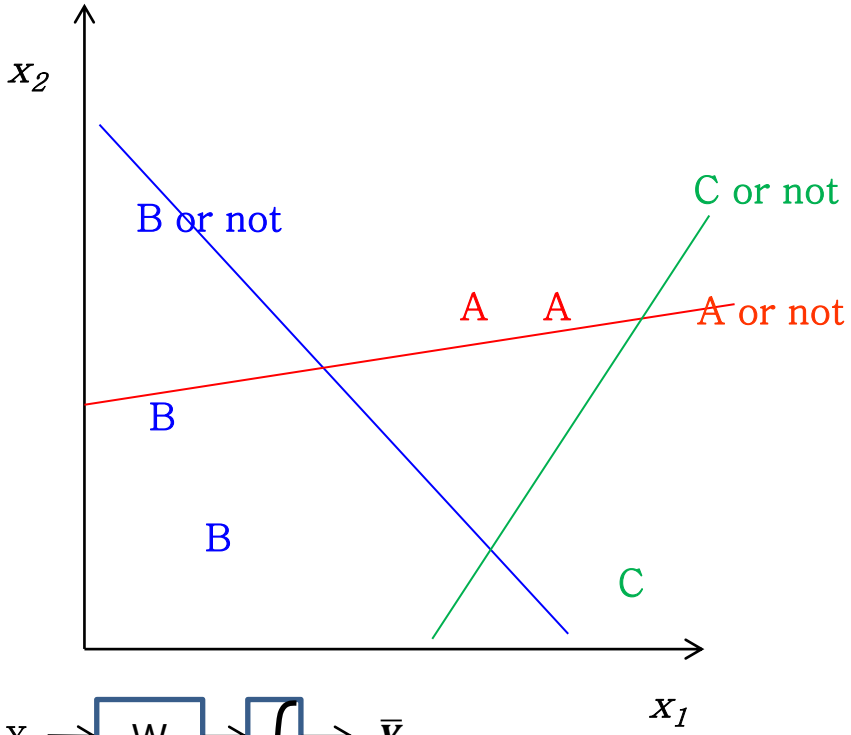
- 결정 공간 (decision surface)

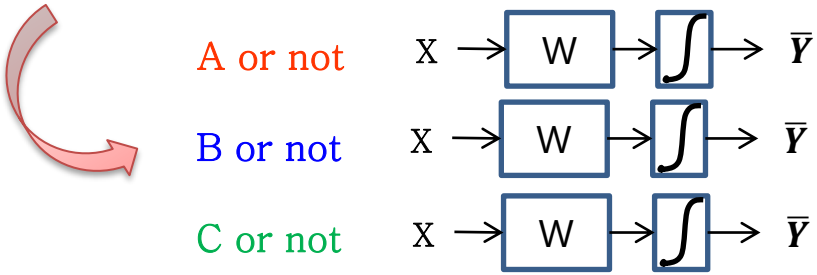


One perceptron → Binominal Classifier

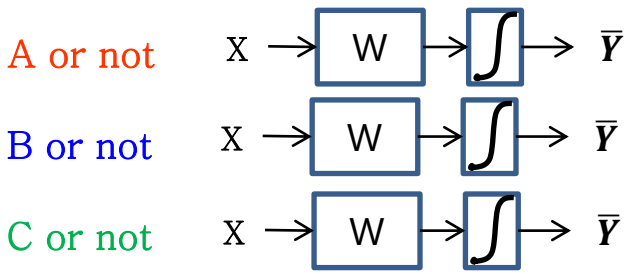
Multinomial Classification

x1 (hours)	x2 (attendance)	y (grade)
10	5	A
9	5	A
3	2	B
2	4	B
11	1	C





Multinomial Classification



$$[w1 \quad w2 \quad w3] \times \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = [w1 \times x1 + w2 \times x2 + w3 \times x3]$$

$$[w1 \quad w2 \quad w3] \times \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = [w1 \times x1 + w2 \times x2 + w3 \times x3]$$

$$[w1 \quad w2 \quad w3] \times \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = [w1 \times x1 + w2 \times x2 + w3 \times x3]$$

$$\begin{bmatrix} w_{A1} & w_{A2} & w_{A3} \\ w_{B1} & w_{B2} & w_{B3} \\ w_{C1} & w_{C2} & w_{C3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \rightarrow \int \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

New Cost Function for Multinomial Classification

Cross Entropy

$S(y) = \bar{Y}$ $\qquad$ $L = Y$

$$\begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \qquad D(S, L) = -\sum_i L_i log(S_i) \qquad \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

[Example]

Y

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

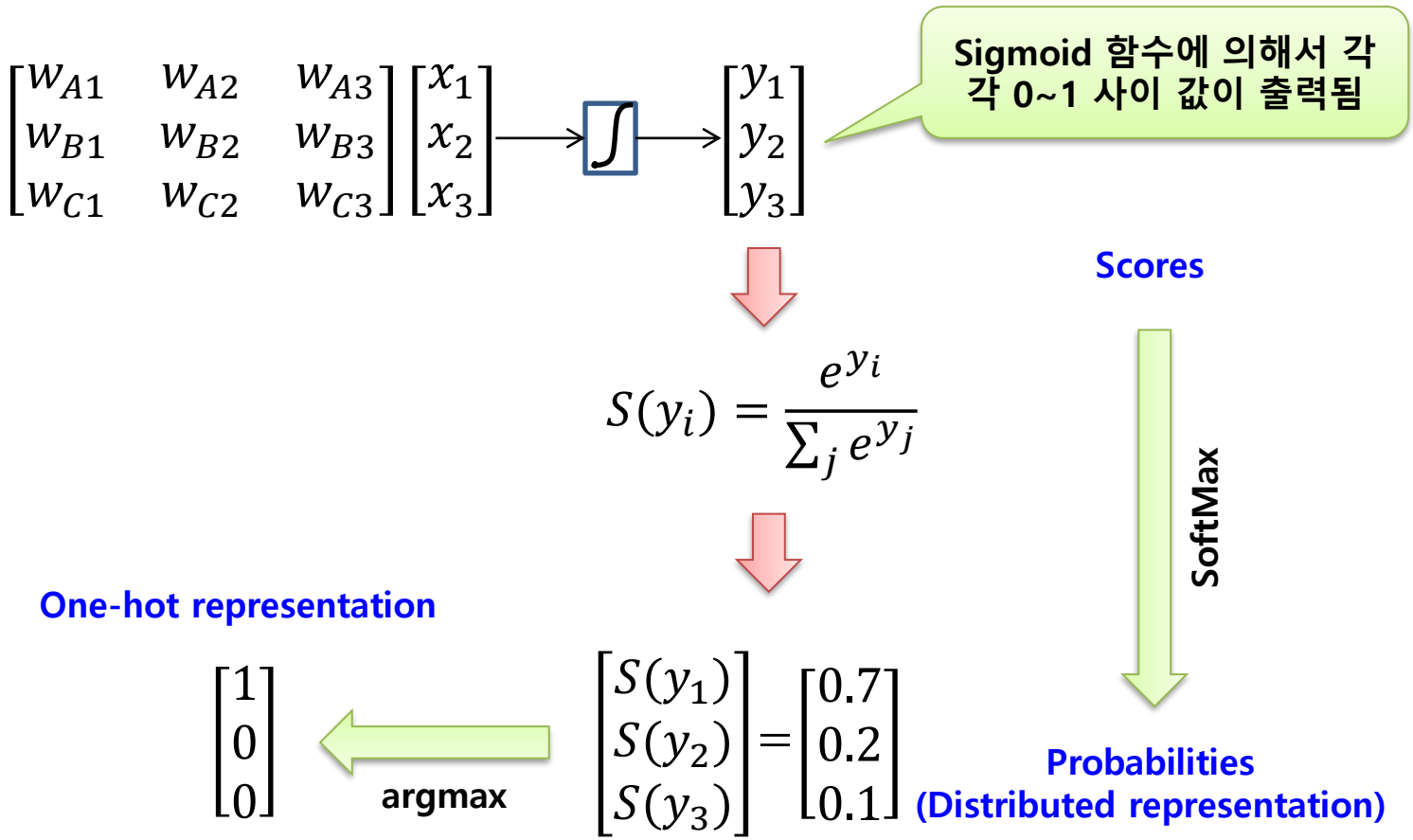
$\bar{Y}$

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \odot -\log\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \odot \begin{bmatrix} \infty \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0$$

$\bar{Y}$

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \odot -\log\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \odot \begin{bmatrix} 0 \\ \infty \end{bmatrix} = \begin{bmatrix} 0 \\ \infty \end{bmatrix} = \infty$$

SoftMax



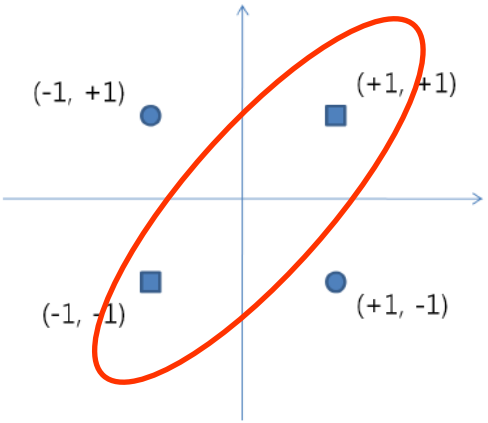
$$\begin{bmatrix} w_{A1} & w_{A2} & w_{A3} \\ w_{B1} & w_{B2} & w_{B3} \\ w_{C1} & w_{C2} & w_{C3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \rightarrow \int \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

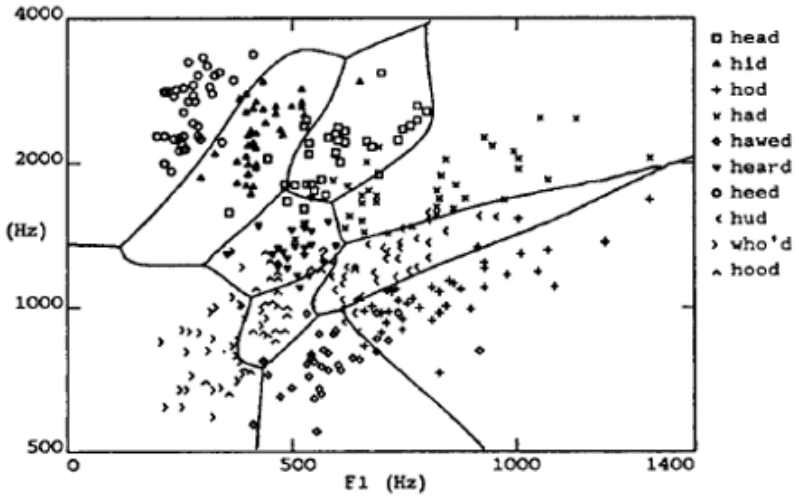
$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

$$\begin{bmatrix} S(y_1) \\ S(y_2) \\ S(y_3) \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Non-linear Problems

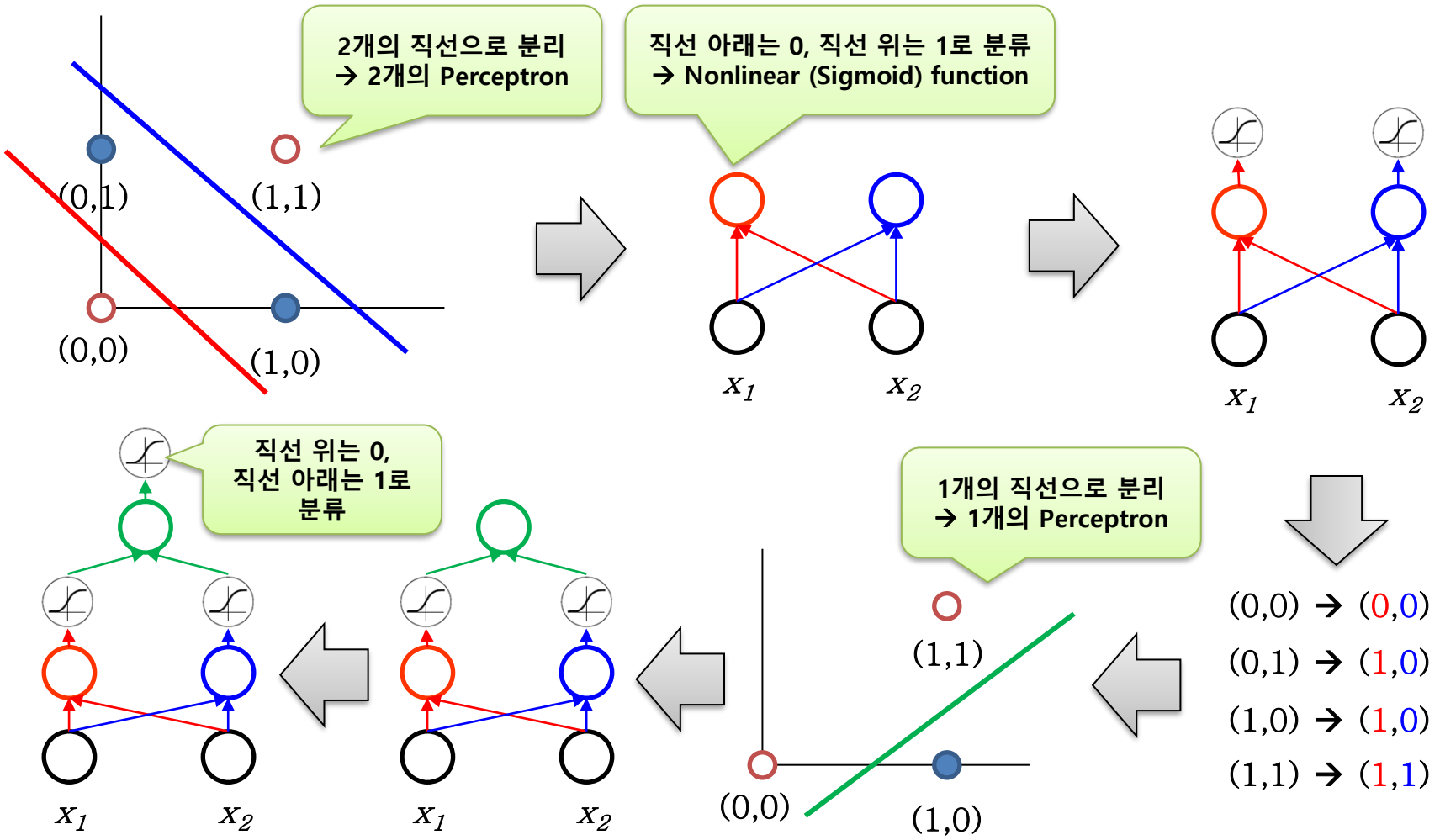
- 비선형 분리 문제





- 비선형 분리 문제 → 선형 분리 문제
  - SVM 커널 함수(kernel function)
  - Single-layer perceptron → Multi-layer perceptron

XOR in Multi-layer Perceptron



질의응답

Homepage: http://nlp.konkuk.ac.kr
E-mail: nlpdrkim@konkuk.ac.kr