

빅데이터와 인공지능 도입

강의자: 김진규
오하이오주립대 박사과정

Contents

I. 빅데이터 혁명: 트렌드

II. 용어 설명

III. 데이터

IV. 머신러닝 기술

V. 지도학습

VI. 비지도학습

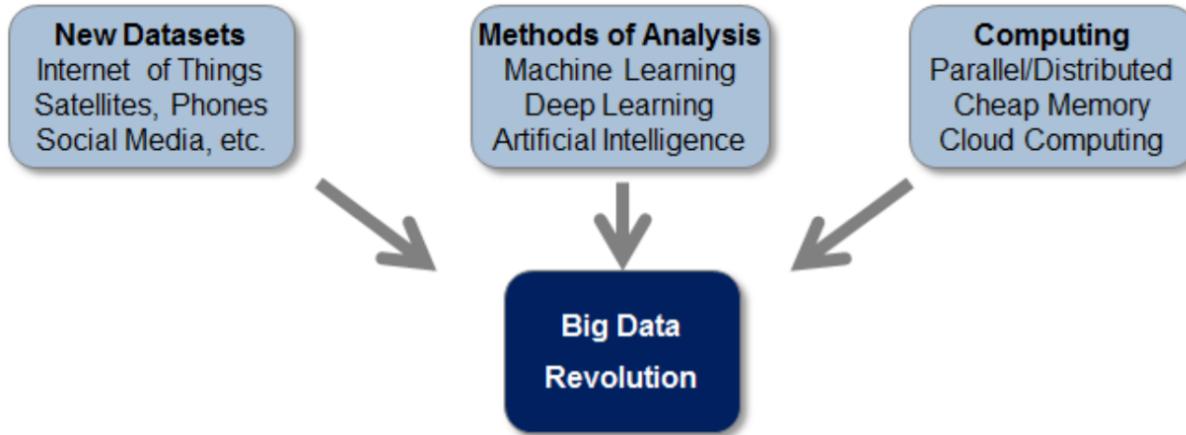
VII. 딥러닝

VIII. 활용

00. Epilogue

1. 빅데이터 혁명: 트렌드

Figure 1: Factors leading to Big Data Revolution



Source: J.P.Morgan Macro QDS.

- 활용 가능한 데이터 양의 폭발적 증가
- 컴퓨터 연산능력과 데이터 저장능력의 향상
- 복잡한 데이터 분석을 위한 머신러닝 방법론의 발전

1. 빅데이터 혁명: 트렌드

□ 데이터 증가

- 현존하는 90%의 데이터가 2년 안에 만들어진 데이터라는 연구결과
- 2015년 4.4 Zb 였던 데이터 양이 2020년에는 44 Zb가 될 것으로 예상
(1 제타바이트 = 1조 기가바이트)
- 촉진 요소
IoT (Internet of Things), 스마트폰, 위성 기술에서의 비용감소

1. 빅데이터 혁명: 트렌드

□ 연산능력과 저장능력: 중요 요소는 “클라우드 컴퓨팅”

- 클라우드 컴퓨팅은 멀리 떨어진 곳에 서버를 두고 여러 사람이 공유된 자원에 접근하는 것

예) 구글에서 웹검색 1회: 약 1000개의 컴퓨터의 협업이 필요함

예: Software-as-a-service (SaaS), Platform-as-a-service (PaaS), etc.

- 결국, 이러한 클라우드 컴퓨팅은 대용량의 데이터를 처리하고 저장하는데에 필요한 시간과 비용을 절약하여 장벽을 줄이는데 기여

1. 빅데이터 혁명: 트렌드

□ 머신러닝 방법론

- 통계학과 컴퓨터과학의 분야
- 구조화되지 않은 큰 데이터를 분석하는데 유용
- 고전적 머신러닝, 딥러닝, 강화학습 등이 그 분야
- 예) 아마존 에코, 구글 홈, 애플 시리

“오직 생산된 데이터의 0.5%만이 현재 분석되고 있다 [Regalado (2013)]”

- 새로운 데이터셋과 머신러닝 방법론을 익히는 것에 대한 설득력 있는 이유

2. 용어 설명

“빅” 데이터

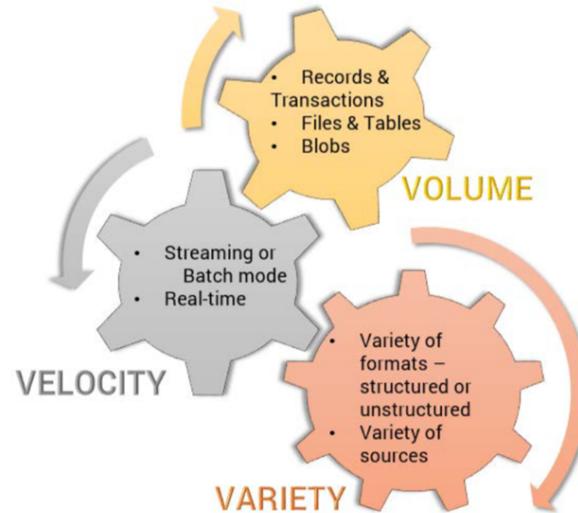
머신러닝

인공지능

2. 용어 설명

□ “빅”데이터: 무엇인 큰 것인가?

Figure 2: Features of Big Data



Source: J.P.Morgan Macro QDS.

- Volume: 데이터의 크기
- Velocity: 데이터를 보내거나 받는 속도가 빠름
- Variety: 데이터의 형태가 다양함 (정형데이터, 비정형데이터 등)

2. 용어 설명

□ 머신러닝

- 목적: 컴퓨터가 특정한 일을 스스로 경험에 의해 학습 가능하게 만드는 것
- 머신러닝의 경우 경험이 누적될 수록 성능이 향상됨
- 분류: 지도학습, 비지도학습, 딥러닝과 강화학습

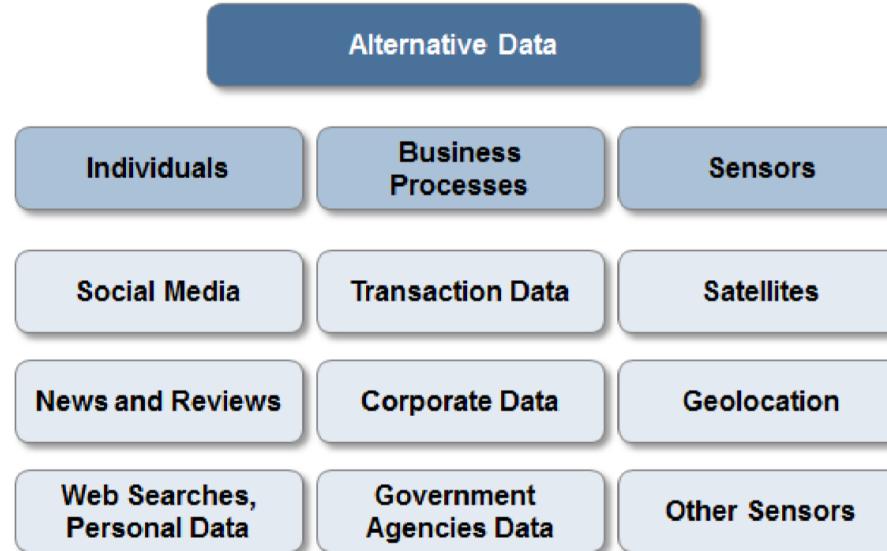
2. 용어 설명

□ 인공지능

- 기계에게 인간과 같은 지능을 갖추게 한다는 넓은 개념
- 처음: 수 많은 규칙을 규정하여 코딩하는 "Symbolic-AI"
- 머신러닝은 인공지능을 얻기 위한 또 다른 하나의 시도
- 머신러닝과 딥러닝은 여태까지 인공지능을 만들기 위한 시도 중 가장 성공적 (이미지 인식, 번역 등)

3. 데이터

Figure 3: Classification of Big/Alternative Data sources



Source: J.P.Morgan Macro QDS.

- 개인에 의해 생성된 데이터
- 비즈니스 프로세스에 의해 생성된 데이터
- 센서에 의해 생성된 데이터

3. 데이터

□ 개인에 의해 생성된 데이터

- 주로 문자 매체에 의해 비정형 데이터로 저장됨

□ 가능한 원천

- 소셜미디어: 트위터, 페이스북, 링크드인 등
- 특별목적 웹사이트: 맛집리뷰사이트 Yelp, 전자상거래 Amazon 리뷰 등
- 웹 검색기록, 개인 기록, 구글검색트렌드 등

3. 데이터

- 비즈니스 프로세스에 의해 생성된 데이터
 - 회사의 비즈니스 프로세스 상에서 형성된 데이터이고 주로 정형데이터
- 예시
 - 신용카드 거래기록
 - 은행거래기록
 - 슈퍼마켓 레코드 기록
 - 공급사슬 데이터

3. 데이터

□ 센서에 의해 생성된 데이터

- 다양한 장치에 부착된 센서에 의해 자동적으로 모아진 데이터이며, 주로 비정형 데이터이며, 데이터 크기가 큰 편임

□ 예시

- 위성사진
- 지정장소에 설치된 카메라 (CCTV등)
- 날씨, 오염 데이터
- IoT (Internet of Things) 등

4. 머신러닝 기술

□ 새로운 방법론의 필요성

- 규모가 큰 비정형데이터의 증가: 스프레드시트 프로그램으로 다루기 어려움
- 따라서, 비정형 데이터를 자동적으로 분석하는 새로운 방법이 필요
- 머신러닝은 몇 가지 놀라운 성과를 이뤄냄
(이미지와 패턴인식, 자연어처리, 자동주행차 등)

“그렇다면, 비즈니스 시장에서 사용되는 머신러닝에는 어떤 종류가 있으며, 각각의 방법론은 어떻게 다른가?”

4. 머신러닝 기술

□ 첫번째 단계: Symbolic AI

- 컴퓨터가 특정한 일을 수행하기 위해 인간이 많은 규칙들을 입력
- 아이디어: 굉장히 많은 규칙들을 컴퓨터에 알려줘서, 컴퓨터가 사람처럼 똑똑하게 보이게 하면 어떨까?
- 그러나, 이러한 경우 컴퓨터는 인간이 수동으로 입력하지 않은, 처음 대면하는 상황에서는 작동하지 않을 것임
- 이것은 기계가 학습한다고 보기 어려움. 수 많은 규칙을 입력한 것
=> 큰 성공이 아님

4. 머신러닝 기술

□ 기계 학습

- 기계학습에서는 컴퓨터가 데이터 인풋(몇 가지 변수들과 데이터셋)과 데이터 아웃풋(인풋 변수의 결과변수)를 입력 받음
- 사람이 기계가 어떤 식으로 학습을 할 지 결정
- 그 이후 기계는 데이터 인풋과 데이터 아웃풋의 관계를 학습
- 3가지 주요 카테고리: 지도학습, 비지도학습, 딥러닝

4. 머신러닝 기술

□ 지도학습

- 정답지가 있음
- 변수를 예측할 수 있는 어떠한 함수를 찾아내는 것이 목표

□ 비지도학습

- 정답지가 없어도 괜찮음
- 데이터 자체의 구조를 학습

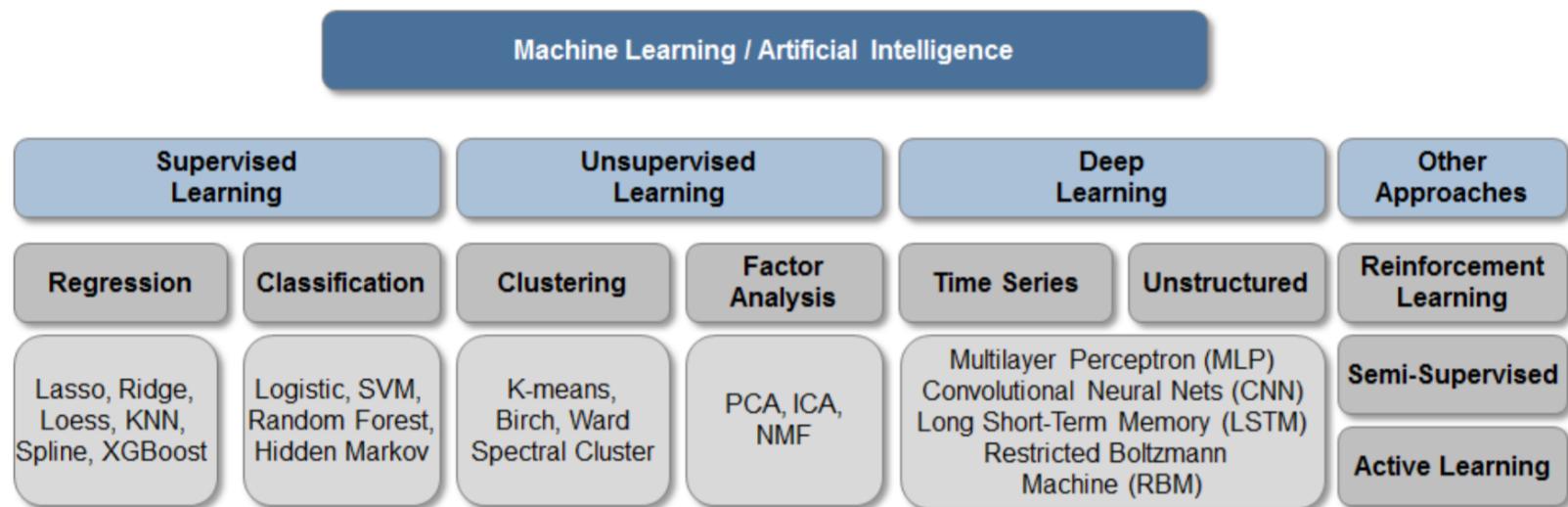
□ 딥러닝

- 인간이 학습하는 과정과 유사함
- 예) 어린 아이가 얼굴을 인식할 때 맨 처음에는 가장 간단한 특징인 눈, 코, 입 인식. 이후 이를 조합해서 더 복잡하고 추상적인 개념인 얼굴을 인식

4. 머신러닝 기술

□ 머신러닝 기술 분류

Figure 7: Classification of Machine Learning techniques



Source: J.P.Morgan Macro QDS

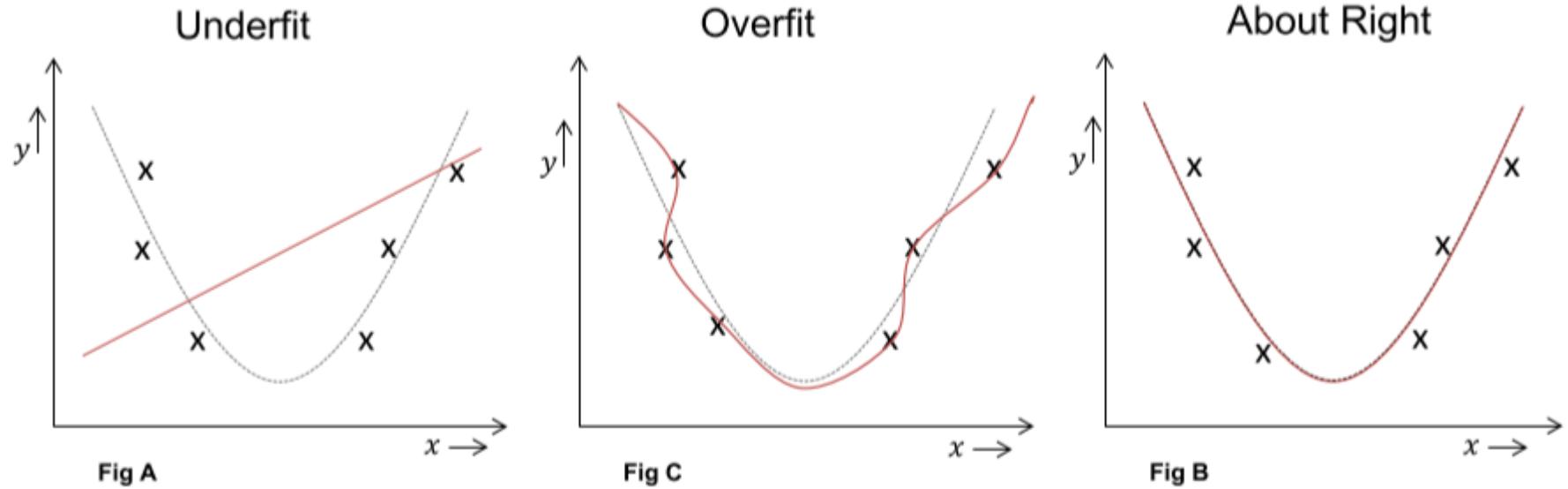
4. 머신러닝 기술

□ 올바른 모델 고르기

- 과적합의 문제

- 정규화 (Regularization): 변수에 적절한 페널티를 부여

Figure 42: Variance-bias trade-off (overfitting vs. underfitting)



Source: J.P.Morgan Macro QDS

5. 지도학습

□ 지도학습

- 알고리즘은 과거의 자료 (인풋, 아웃풋 변수)를 받아 학습을 통해 샘플 밖의 데이터를 가장 잘 예측하는 것을 목표로 함

□ 회귀분석과 분류

- 회귀분석: 몇 가지 인풋변수를 기반으로 아웃풋 변수를 예측
(주가 수익률 예측)
- 분류: 아웃풋 변수를 카테고리로 묶어 분류
(내일 팔 것인지, 살 것인지)

5. 지도학습

□ 선형 회귀분석

- 단순 선형 회귀모형 또한 지도학습의 일종으로 생각할 수 있음
- 하지만 종종 선형회귀모형은 실제 관계를 대표하기에는 너무 간단한 모형일 수 있음

$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i + \varepsilon$$

OLS: Minimize Historical Sum of $\left(y - (\beta_0 + \sum_{i=1}^n \beta_i x_i) \right)^2$.

5. 지도학습

□ 페널티 회귀분석

$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i + \varepsilon$$

OLS: Minimize Historical Sum of $(y - (\beta_0 + \sum_{i=1}^n \beta_i x_i))^2$.

Lasso: Minimize Historical Sum of $(y - (\beta_0 + \sum_{i=1}^n \beta_i x_i))^2 + \alpha \sum_{i=1}^n |\beta_i|$,

Ridge: Minimize Historical Sum of $(y - (\beta_0 + \sum_{i=1}^n \beta_i x_i))^2 + \alpha \sum_{i=1}^n \beta_i^2$,

Elastic Net: Minimize Historical Sum of $(y - (\beta_0 + \sum_{i=1}^n \beta_i x_i))^2 + \alpha_1 \sum_{i=1}^n |\beta_i| + \alpha_2 \sum_{i=1}^n \beta_i^2$.

- 불필요한 베타와 아주 큰 베타를 영으로 만드는 회귀

5. 지도학습

K-Nearest Neighbor (KNN)

- 새로운 데이터가 주어졌을 때, 과거 데이터와 비교해서, 과거의 어떤 상황과 가장 유사했는지, 가장 유사한 K개의 사례를 추출
- K개의 가장 가까운 이웃을 찾은 뒤 그것을 평균해서 우리의 예측값으로 사용

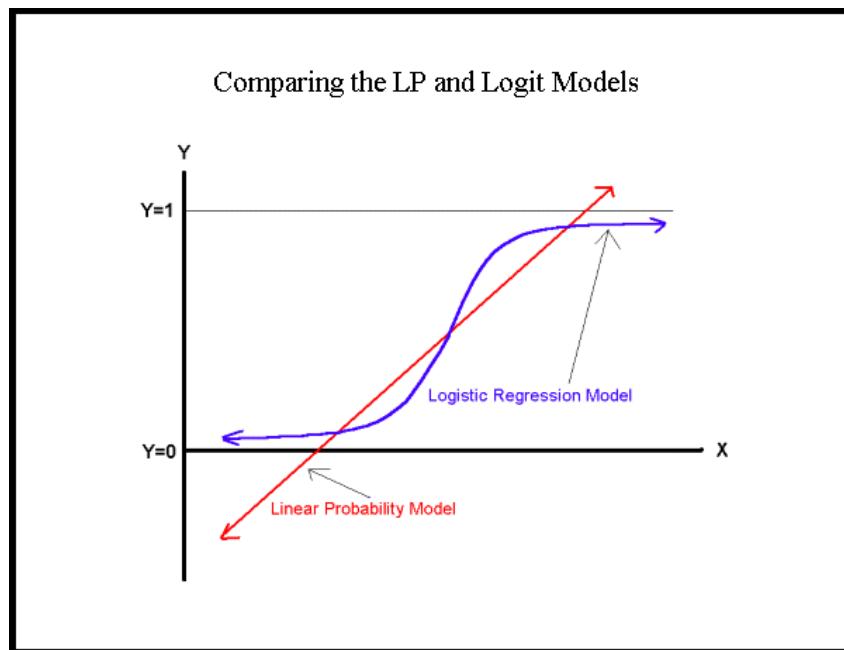
LOESS

- K개의 가장 가까운 이웃을 고른 뒤, 평균이 아니라 선형 회귀분석을 해서 그 예측 값을 사용

5. 지도학습

□ Logistic Regression

- 2개로 구성된 예측값을 생성 (예를 들면 buy or sell)
- 과거 데이터를 바탕으로 해당 이벤트의 발생 확률을 예측

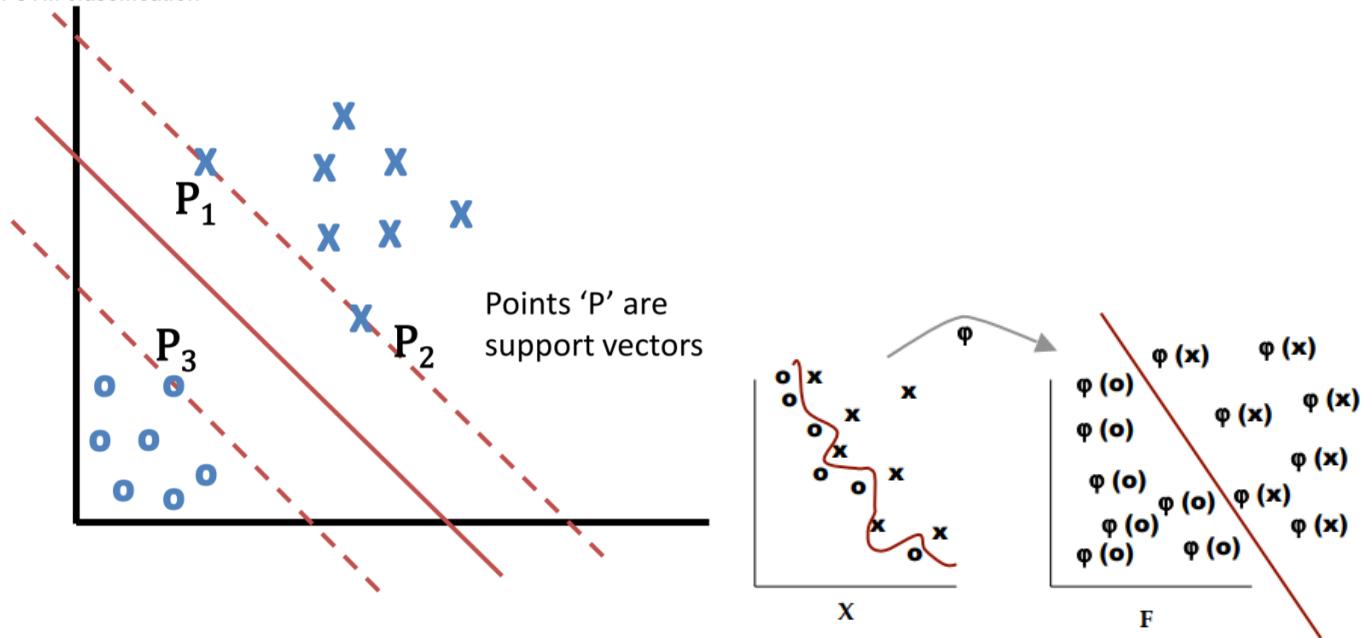


5. 지도학습

□ Support Vector Machine (SVM)

- 가장 유명한 분류 알고리즘
- 특징을 추출하고, 축을 변형하고, 더 높은 차원을 이용해 보강

Figure 58: Simplified illustration of SVM classification



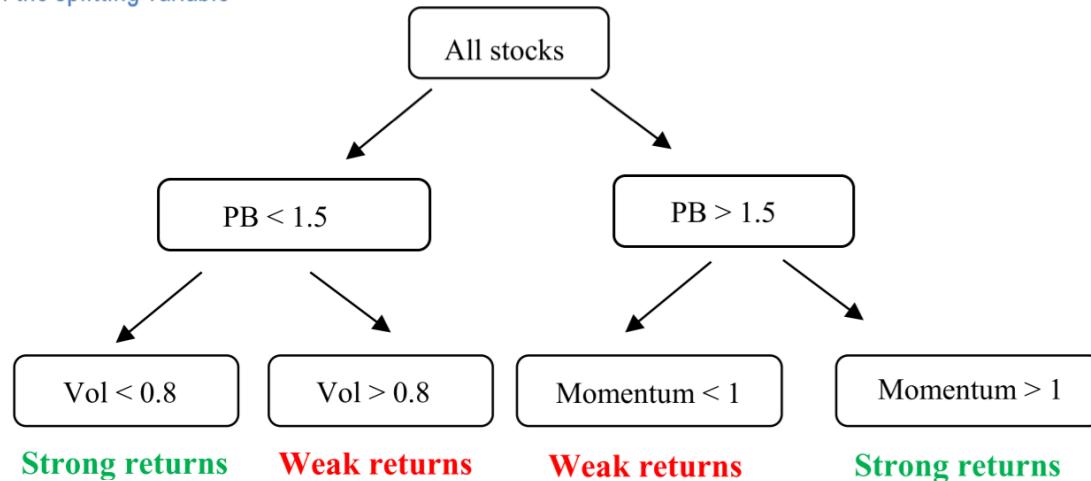
Source: J.P.Morgan Macro QDS

5. 지도학습

□ 의사결정나무

- 본질적으로 경영 의사결정에서 사용되는 플로우차트와 유사
- 의사결정나무를 모델링하기 위해서 알고리즘은 모든 가능성 중에 가장 최적의 나누는 값을 찾고 가지의 수를 찾음

Figure 63: Example of decision tree to classify whether future stock returns are good or poor. In each node, we split the data into two subsets depending on the value of the splitting variable



5. 지도학습

□ 랜덤포레스트

- 하나의 의사결정나무는 샘플 내 데이터의 구조를 매우 잘 구조화 하지만, 샘플 밖의 데이터에서는 잘 작동하지 않을 수 있음
 - 이때 배깅을 사용하게 되는데, 이는 원래 데이터에서 많은 랜덤 샘플을 뽑아내는 것을 의미
 - 그렇게 뽑아낸 랜덤 샘플에서 각각을 의사결정나무로 추정
 - 그렇게 추정한 의사결정나무를 평균
- => 예측치의 변화를 줄이게 됨

6. 비지도학습

□ 비지도학습

- 변수들 사이의 관계 혹은 그것들의 공통 요인을 찾음
- 비지도학습에서 기계는 단순히 전체 데이터를 입력 받고, 무엇이 독립변수인지 무엇이 독립변수의 결과로 나온 변수인지 입력 받지 않음

□ 클러스터링과 요소분석

- 클러스터링: 데이터를 유사성을 기반으로 더 작은 그룹으로 나눔
- 요소분석: 데이터를 이끄는 주요 요소가 무엇인지 분석

6. 비지도학습

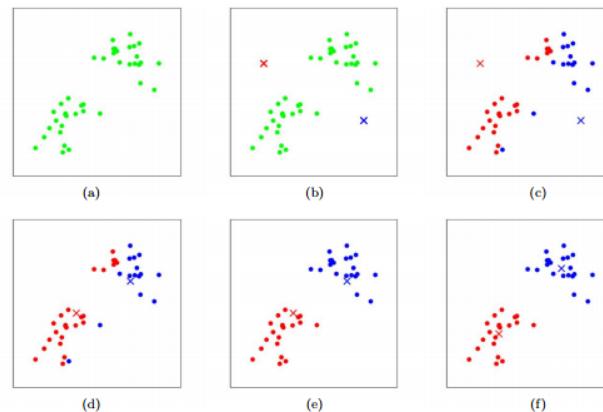
□ 클러스터링

- 비지도학습에서 알고리즘은 에러 제곱항의 합을 최소화 하려고 함.
- 즉, $(\text{predicted value} - \text{actual value})^2$ 를 최소화하려고 함
- 하지만 비지도학습에서는 실제 값을 모름
- 이러한 상황에서 어떻게 에러를 계산해야 할까?

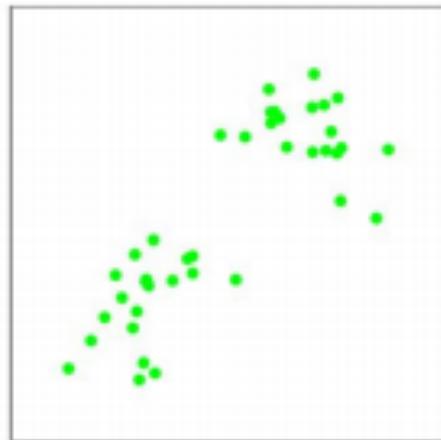
6. 비지도학습

□ K-means clustering

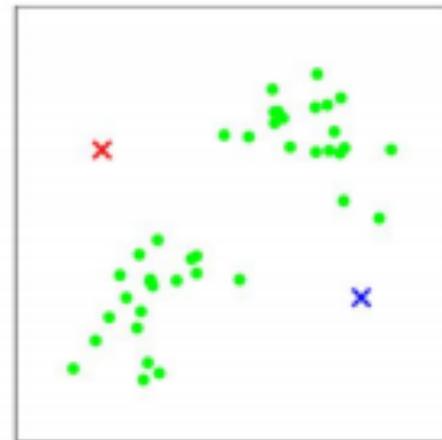
- 가장 간단한 클러스터링 알고리즘
- 알고리즘은 맨 처음 랜덤한 하나의 점을 중심점으로 잡아 시작
- 반복적으로 다음 두 단계를 시행하여 수렴할 때까지 진행
 1. 각각의 점을 가장 가까운 중심점에 연결
 2. 중심점을 속한 모든 점들의 평균으로 재설정



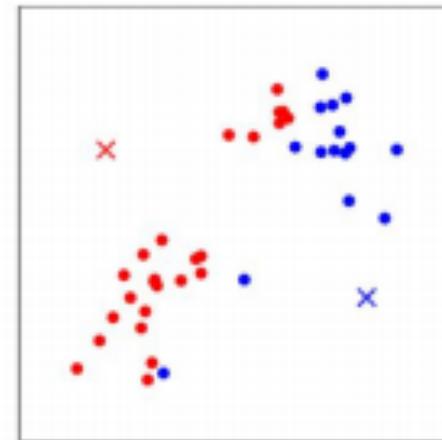
6. 비지도학습



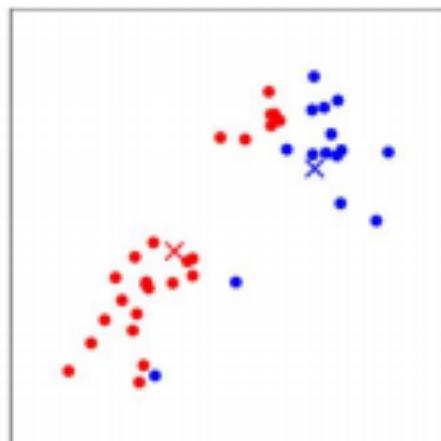
(a)



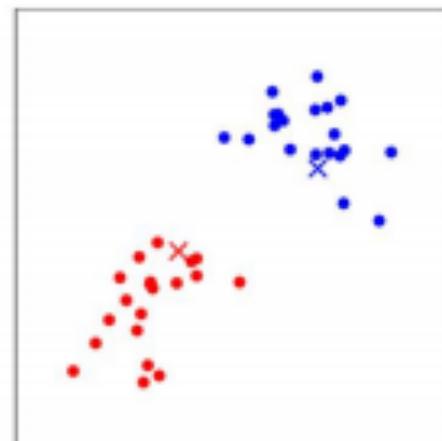
(b)



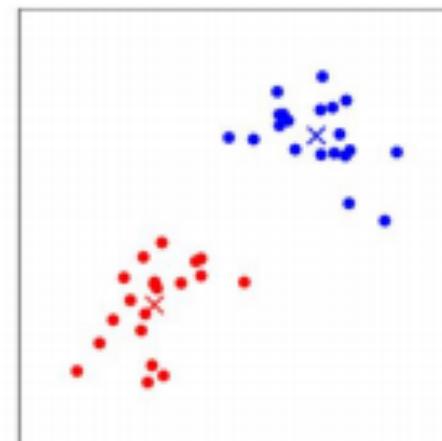
(c)



(d)



(e)



(f)

6. 비지도학습

요소분석

- 목표: 데이터에 가장 영향을 주는 요소를 찾고 이에 따라 관측치를 분류

주성분분석

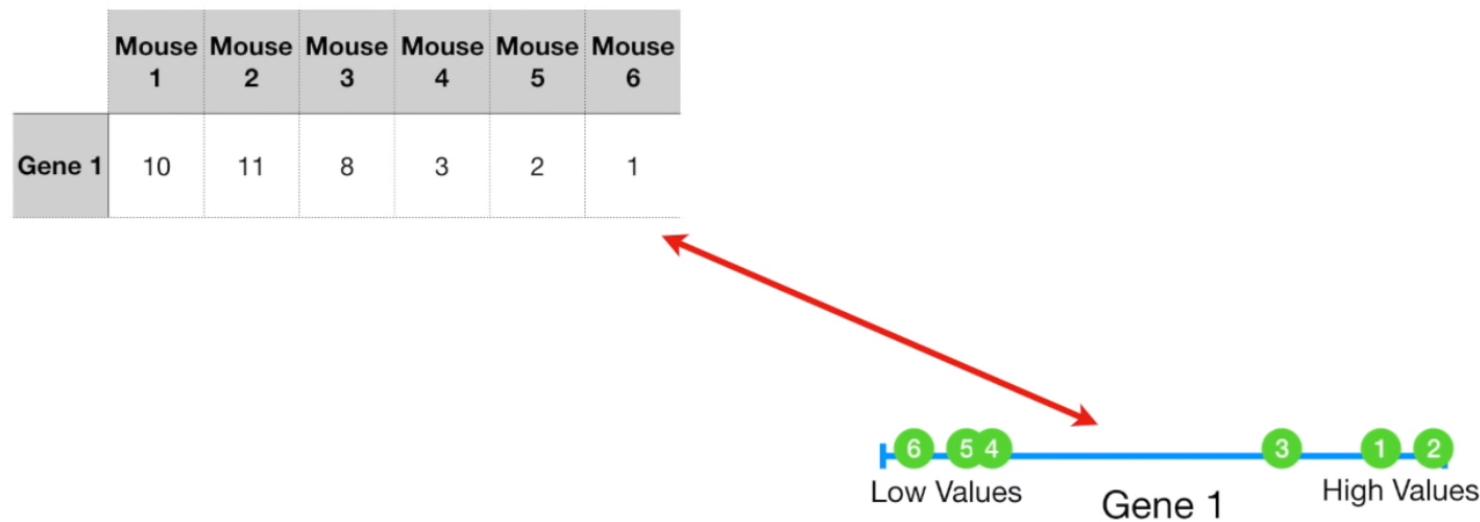
- 요소분석에서 아주 잘 알려진 분석

- 원래 차원축소에 이용되던 통계적 방법론

6. 비지도학습

□ 주성분분석

- 유사성을 기반으로 분류를 하고 싶다고 생각해보자
- 만약 하나의 특징변수만을 가지고 있다면 그 특징변수를 기준으로 각각을 분류할 수 있음

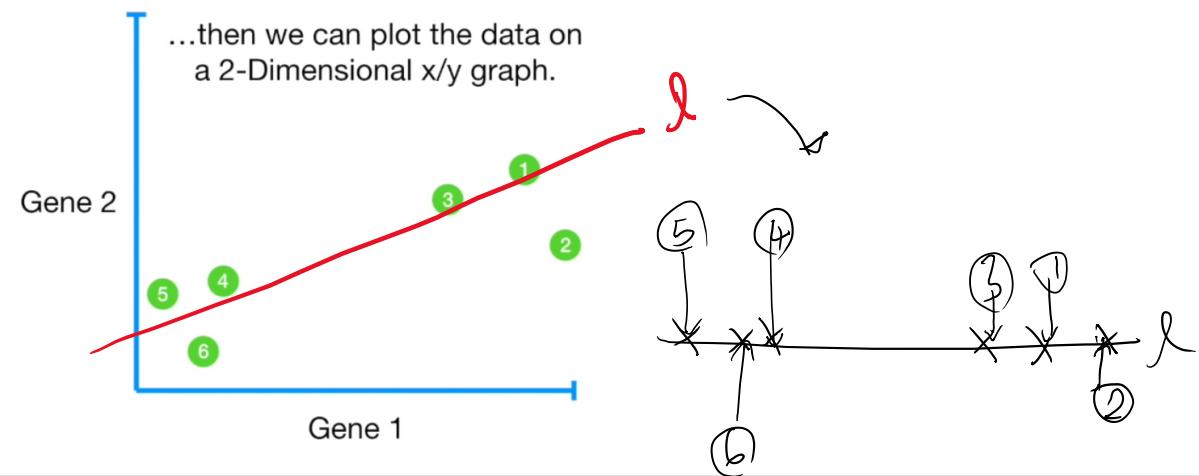


6. 비지도학습

□ 주성분분석

- 만약 두 가지 특징변수가 있다면, 두 개의 축을 이용해 분석 가능
- 또한, 하나의 선분 I 을 이용해 비교가능
- 이 하나의 선분이 주성분의 개념
- 만약 n개의 특징변수가 있다면 어떻게 분석해야 할까?

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

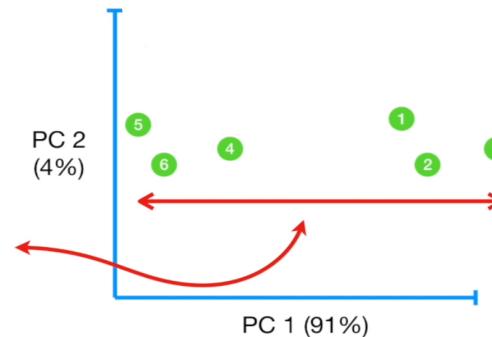


6. 비지도학습

□ 주성분분석

- 주성분분석은 데이터를 분류하는데 가장 중요한 특징을 가지고 있는 변수 혹은 변수들의 조합을 찾아내는 기술 (가장 variation을 크게 만드는 변수)
- 즉, 주성분은 전체 샘플의 차이를 설명하는 가장 중요한 variation이라고 생각할 수 있음
- 이러한 특징을 찾아낸 이후, 이 가장 중요한 특징을 기준으로 차이를 분석하게 되면 실제 n개의 변수가 있어도, 가장 중요한 k개의 변수를 기준으로 분석할 수 있음

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7



7. 딥러닝

□ 딥러닝

- 여러 층에 걸쳐서 학습
- 가장 간단한 컨셉을 학습한 후 이를 조합하여 더 복잡하고 추상적인 개념을 만들어냄

□ 딥러닝과 고전적 머신러닝

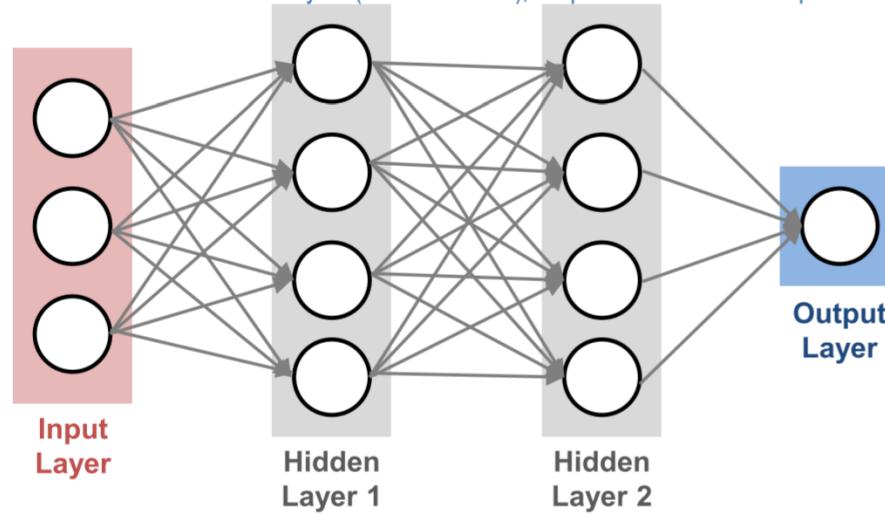
- 고전적 머신러닝은 수치적 예측과 분석에서 탁월
- 그러나 고전적 머신러닝은 직관적 일을 해내려고 고안된 것이 아니고, 잘 하지도 못함 (이미지를 인식하고 텍스트를 분석하는 것 등)
- 딥러닝의 목표는 사람이 정의하기는 어렵지만 행하기는 쉬운 일들을 해내는 알고리즘을 만드는 것

7. 딥러닝

□ 작동원리

- 뉴런: 아주 단순한 계산기라고 생각 가능. 계산기가 하는 일은 첫째, 인풋으로 들어온 값들을 가중평균하고, 둘째, 만약 가중평균 값이 미리 정해놓은 어떤값을 초과하면 다른 뉴런에게 전달

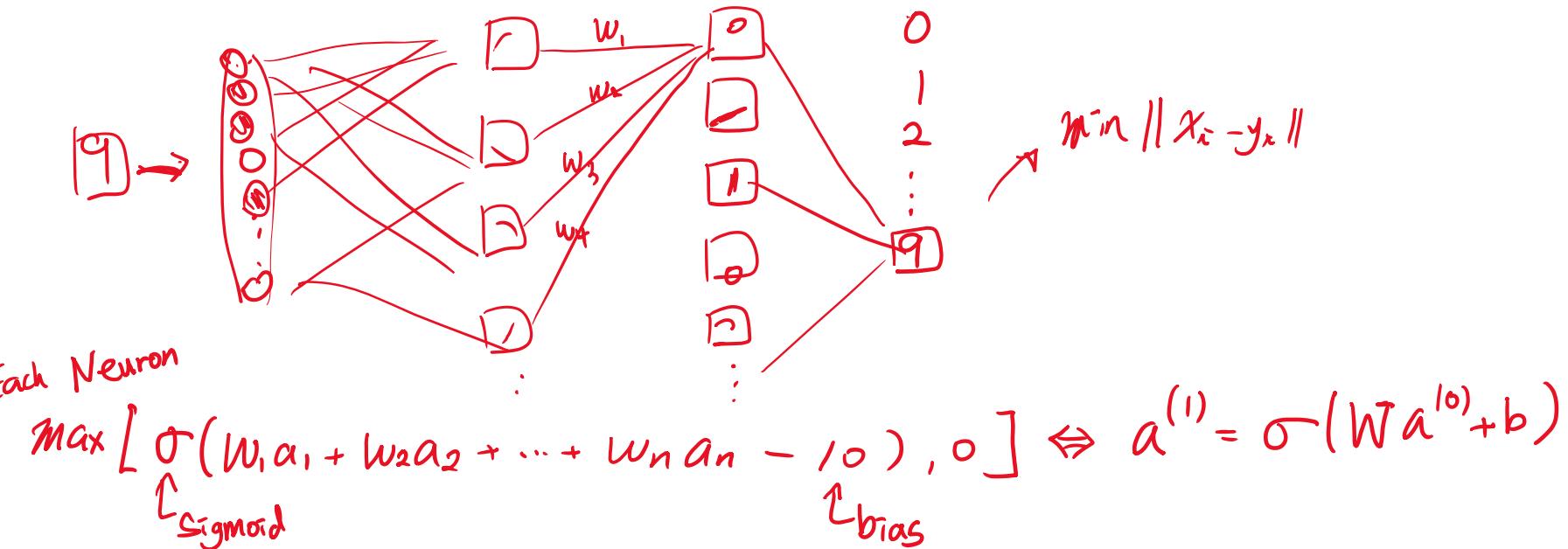
Figure 85: Example of a Neural Network with 2 hidden layers (4 neurons each), 3 input variables and 1 output variable



Source: J.P.Morgan Macro QDS

7. 딥러닝

□ 예제: 손글씨 인식



- 학습이란 적절한 가중치와 바이어스를 찾는 것, i.e., W & b
- 이를 가장 단순한 형태의 MLP (Multilayer Perceptron)라고 함

7. 딥러닝

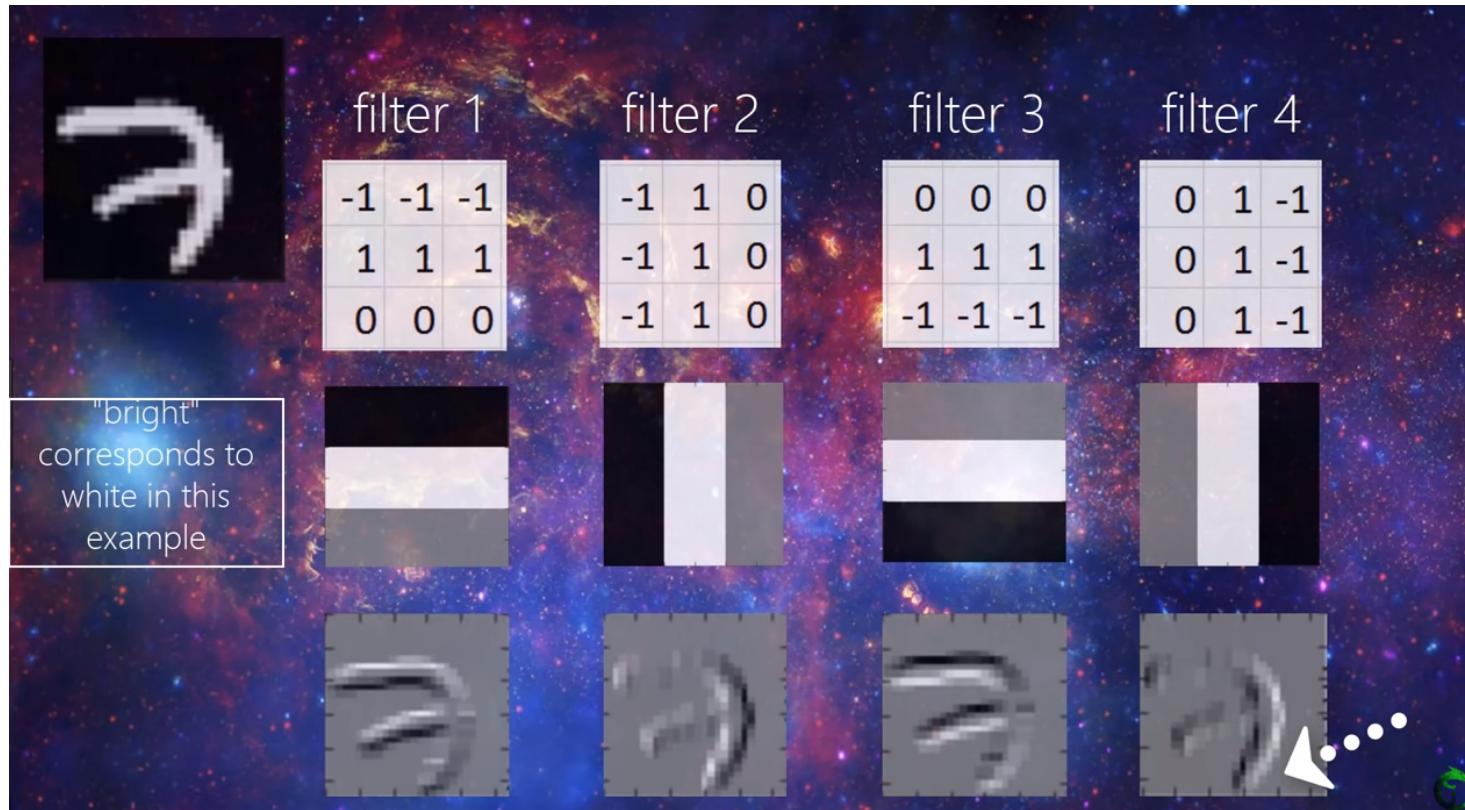
CNN (Convolutional Neural Network): Extension1

- 컨벌루셔널 필터를 사용해 이미지의 엣지 등을 잘 인식하도록 하는 방법

7. 딥러닝

- CNN (Convolutional Neural Network): Extension1

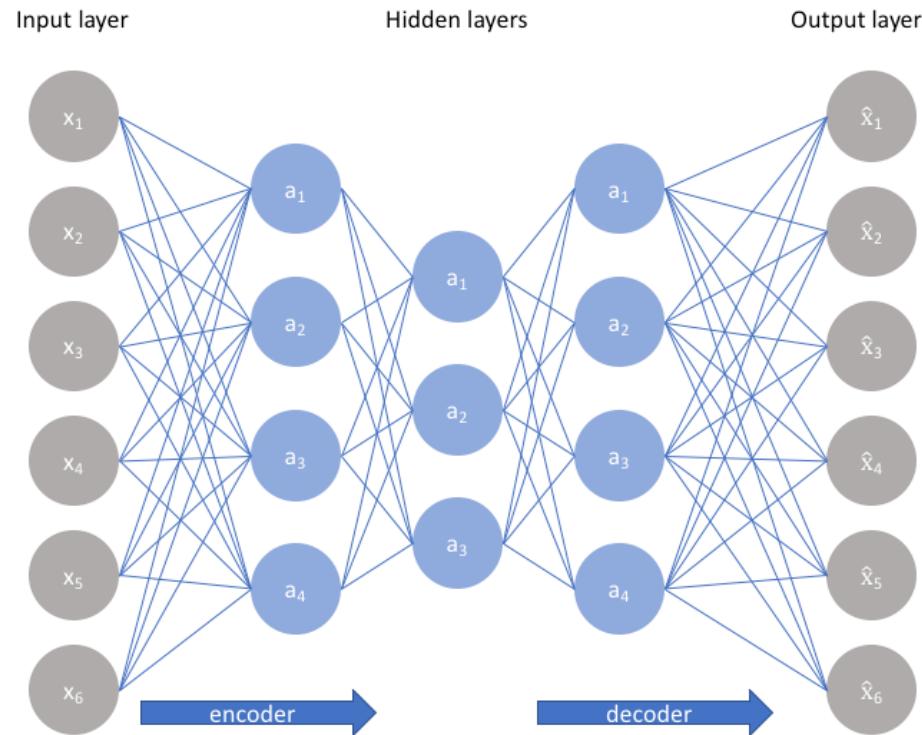
- 하나의 단순한 컨벌루셔널 필터 예



7. 딥러닝

□ Auto-encoder: Extension2

- 우리가 추출한 특징을 리버스 엔지니어링을 해서 원래값과 비교를 하면 어떨까?



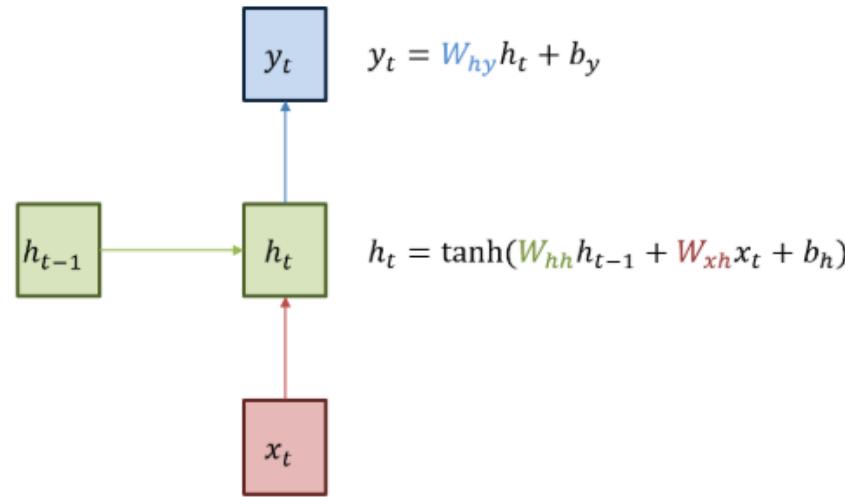
7. 딥러닝

□ RNN (Recurrent Neural Network): Extension3

- 과거의 문맥을 학습모델에 포함시키면 어떨까?

예) 다음에 올 단어를 예측하는 것 h,e,l,l,o

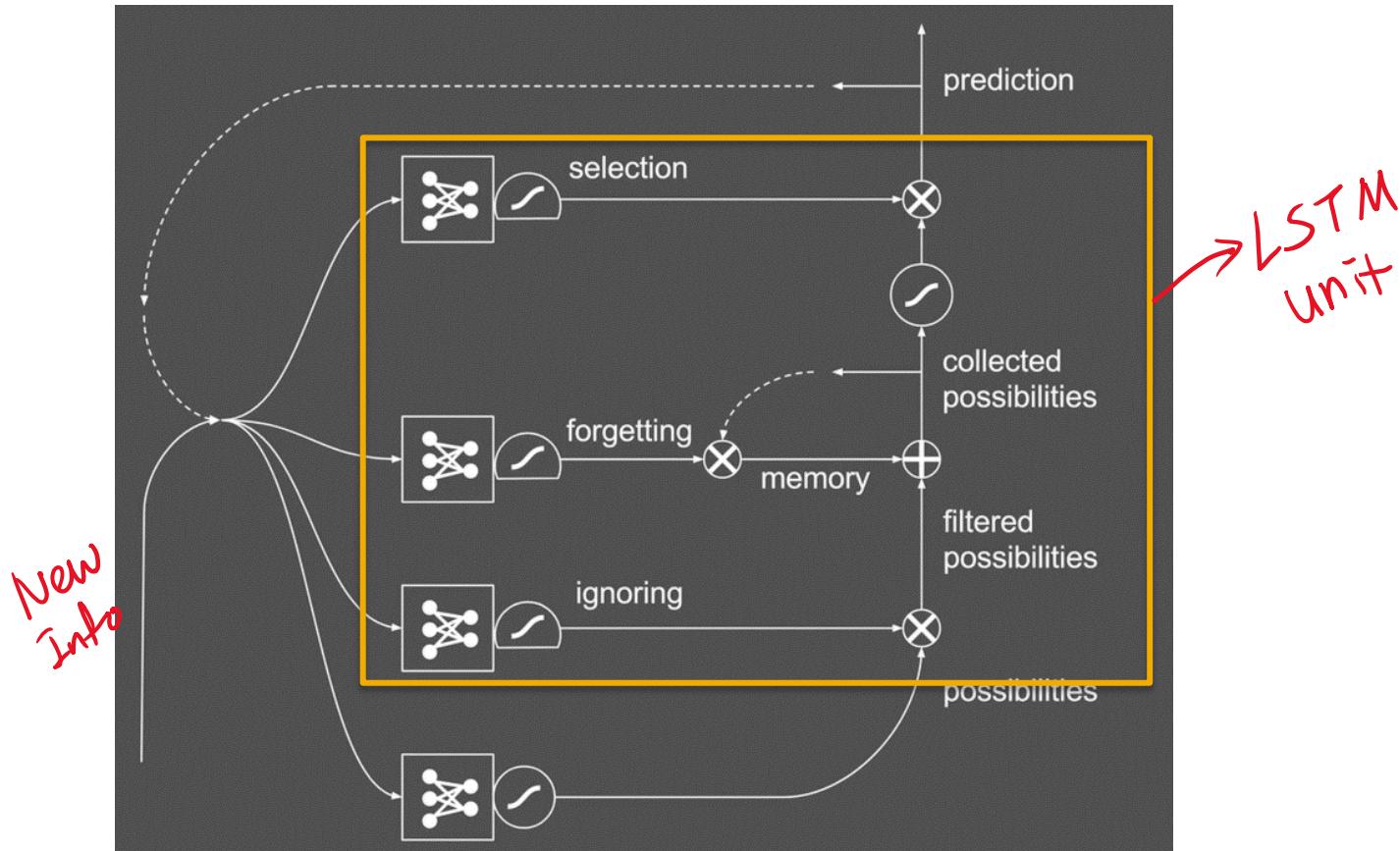
- 그래디언트 소실문제: 장기기억을 불가능하게 만듬
(곱의 개념, tanh의 함수 형태)



7. 딥러닝

□ LSTM (Long-Short Term Memory): Extension4

- RNN에 특별한 박스를 더해서 그라디언트 소실문제를 해결



7. 딥러닝

- LSTM 예제: 외국 어린이 책을 만들어보자

- 요소: Jane, Spot, Doug, saw, .

- 목표:

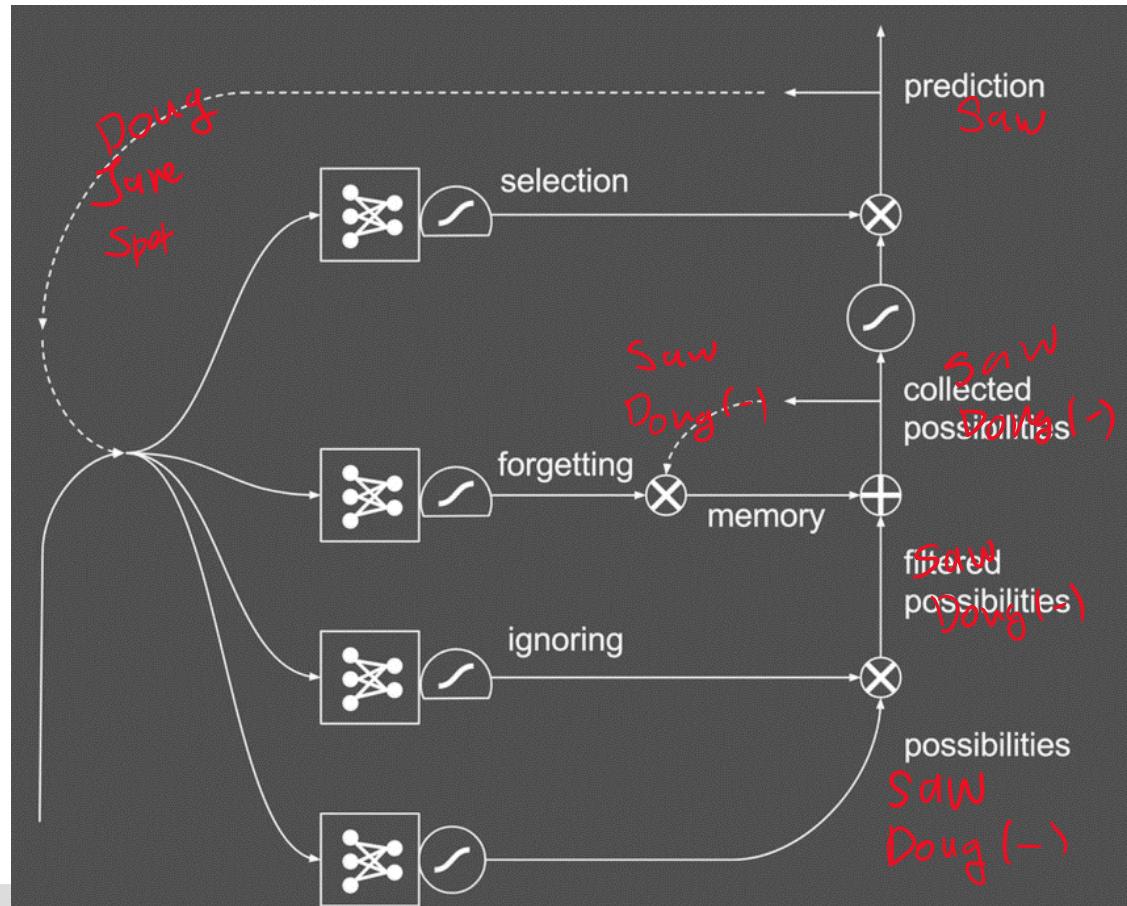
Jane saw Spot.
Spot saw Doug.
Doug saw Spot. ...

- 가능한 실수:

Jane saw Jane.
Doug saw Spot saw ...
Doug.

- 현재:

Jane saw Spot.
Doug ...



7. 딥러닝

- LSTM 예제: 외국 어린이 책을 만들어보자

- 요소: Jane, Spot, Doug, saw, .

- 목표:

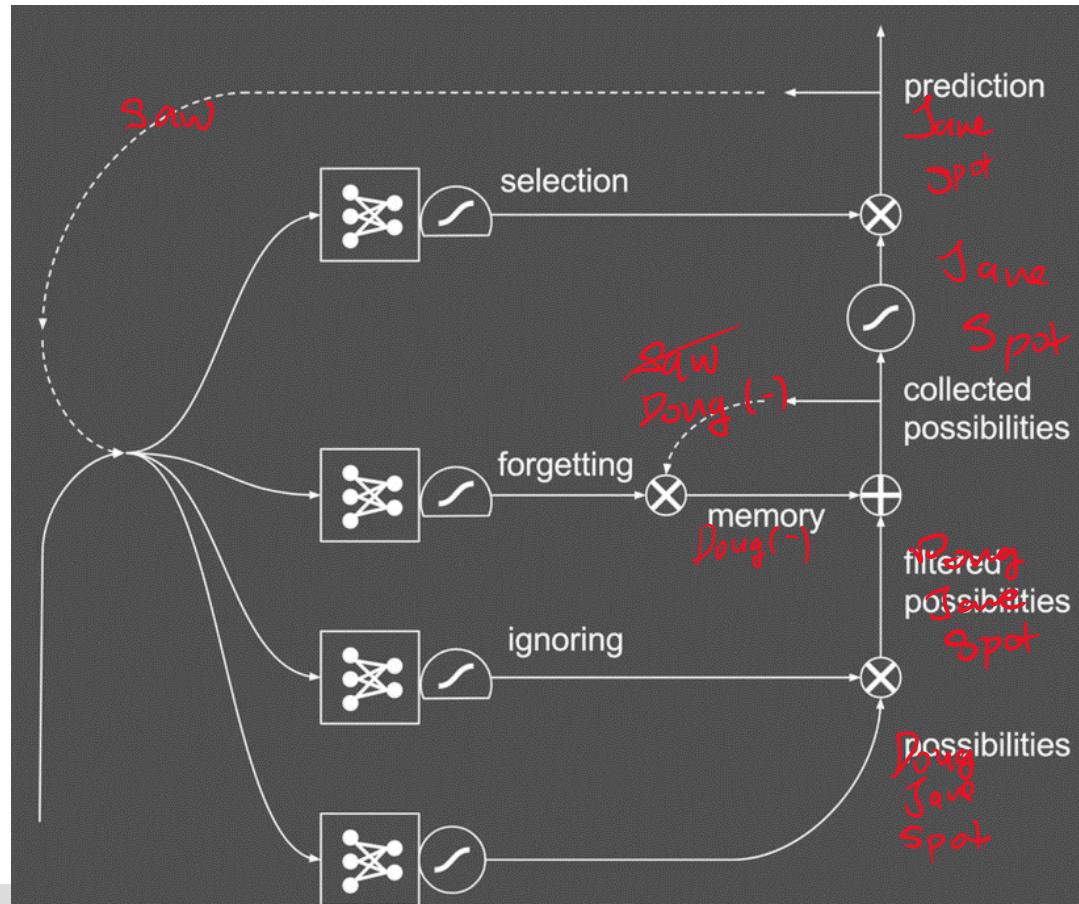
Jane saw Spot.
Spot saw Doug.
Doug saw Spot. ...

- 가능한 실수:

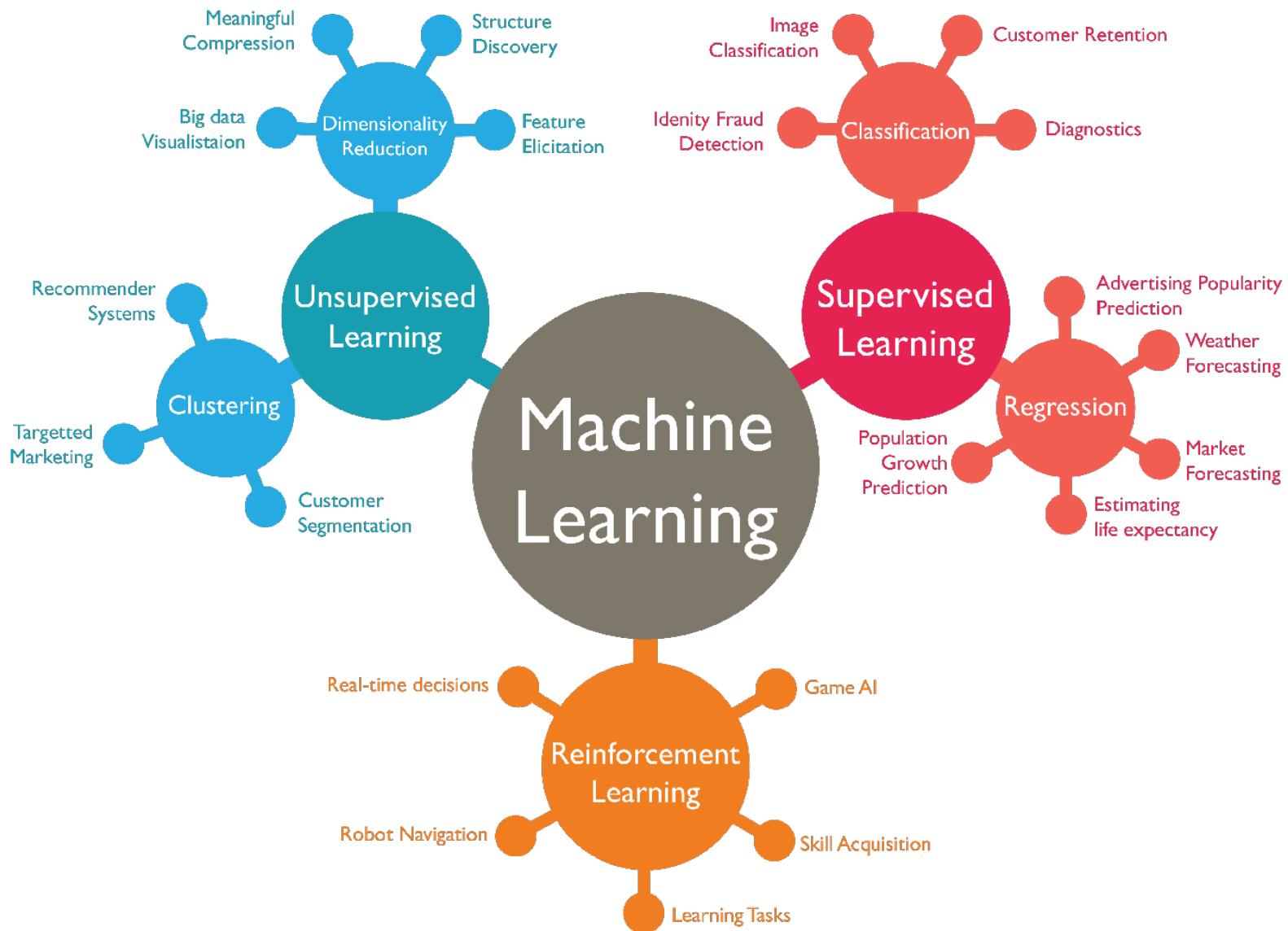
Jane saw Jane.
Doug saw Spot saw ...
Doug.

- 현재:

Jane saw Spot.
Doug saw...

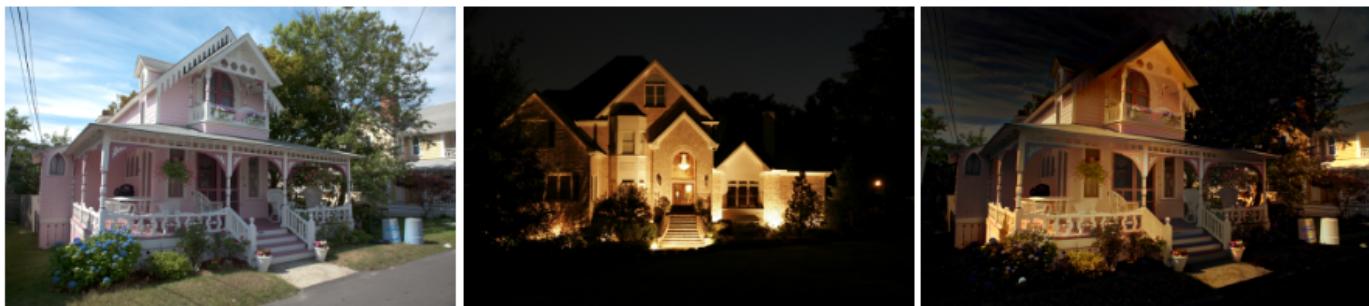


8. 머신러닝 활용



8. 머신러닝 활용

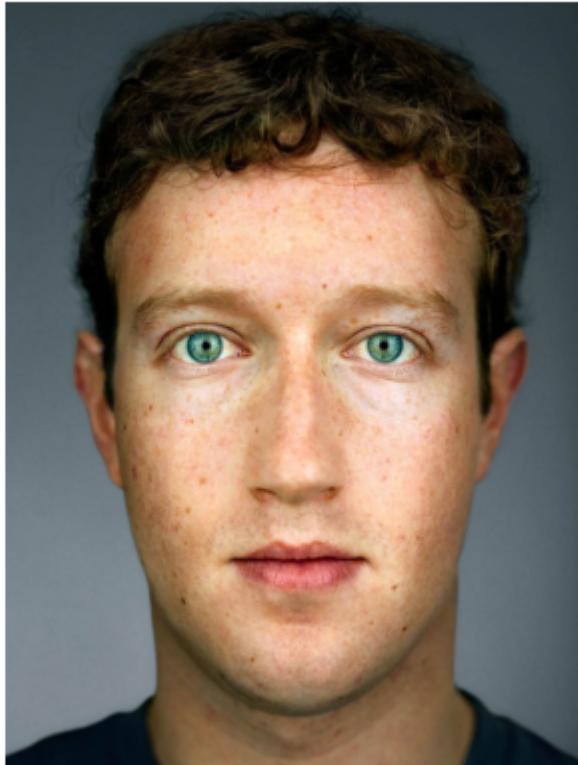
- deep-photo-styletransfer: input, style, output



8. 머신러닝 활용

- deep-photo-styletransfer: input, style, output

- 스타일과 원래 컨텐츠를 어떻게 매핑할까?
- CNN의 확장



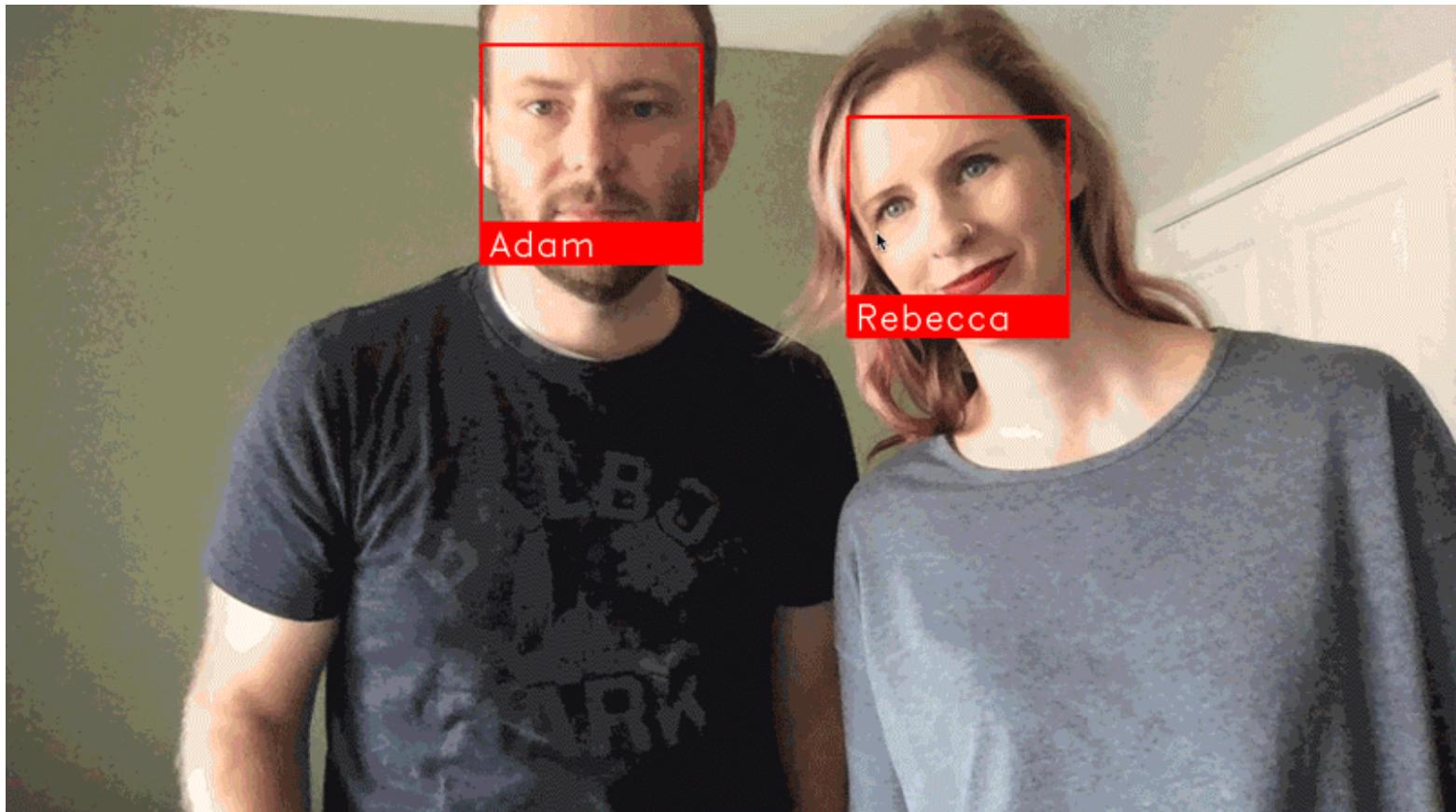
8. 머신러닝 활용

- deep-photo-styletransfer extension



8. 머신러닝 활용

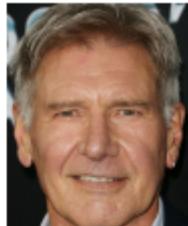
□ 얼굴인식



8. 머신러닝 활용

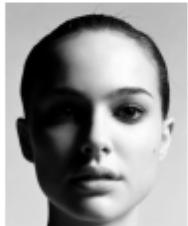
□ 얼굴인식

- 극복해야할 요소



(a)

(b)



(c)



(d)

8. 머신러닝 활용

□ 얼굴인식

- 얼굴 파악: 얼굴의 위치를 파악
- 얼굴 조정: 크기를 맞추고 잘라서 원래 레퍼런스 이미지에 맞게 조정
- 얼굴 대표: 얼굴 사진의 픽셀값을 특징벡터로 변형
- 얼굴 매칭: 특징벡터와 매칭되는 유사한 얼굴을 찾음
- CNN 사용



Fig. 2: Face recognition building blocks.

8. 머신러닝 활용

- Magenta: Art and Music composer

- RNN



8. 머신러닝 활용

- Google: Neutral Networks and 'Machine that Dream'



구글의 뉴럴넷에서 가져온 이미지

8. 머신러닝 활용

- Deep Voice Conversion: Neural Voice Cloning



8. 머신러닝 활용

□ 야라 인터네셔널

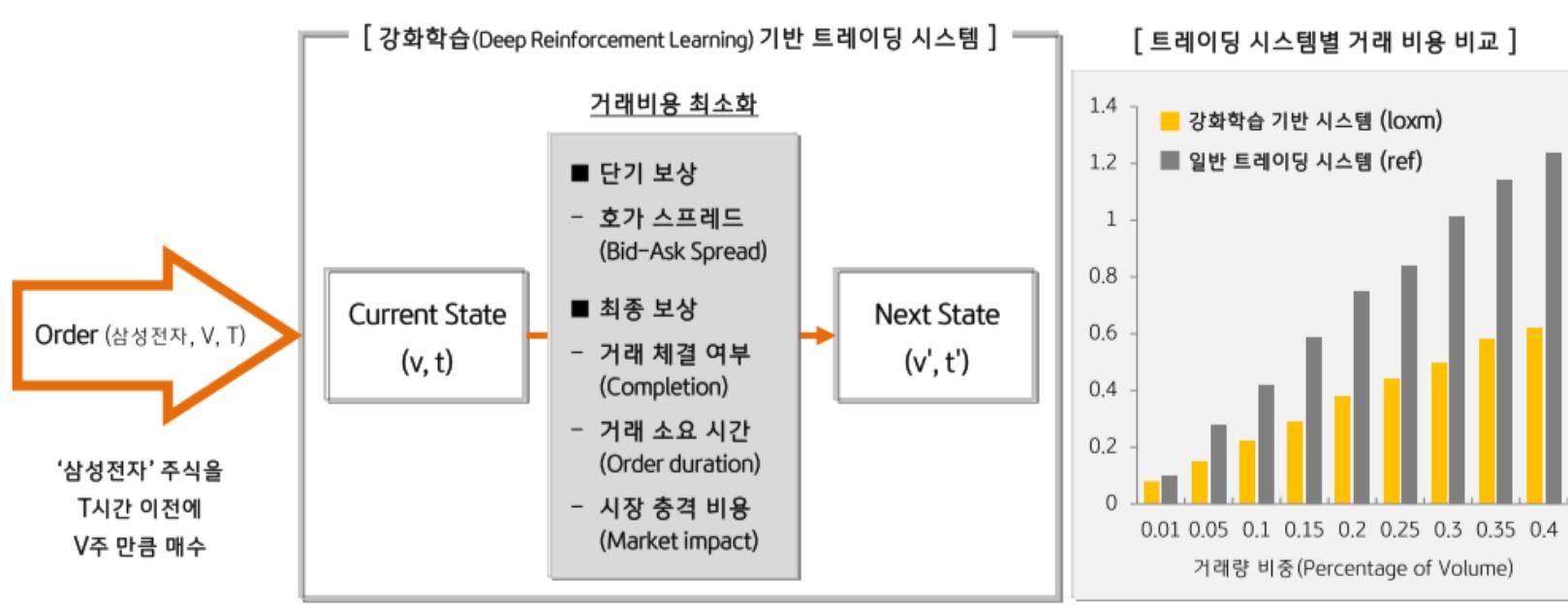
- 세계 최대 비료회사이자 환경솔루션 공급업체 + IBM
- 각 농지와 작물 요구에 맞춰 실시간으로 권고사항 제시



8. 머신러닝 활용

□ LOXM: JP Morgan Trading system

- 호가스프레드, 거래체결여부, 거래소요시간, 시장충격비용으로 비용함수를 만든 뒤, 이를 최소화 하는 방법을 사용
- 벤치마크 시스템과 비교하여 훌륭한 성과

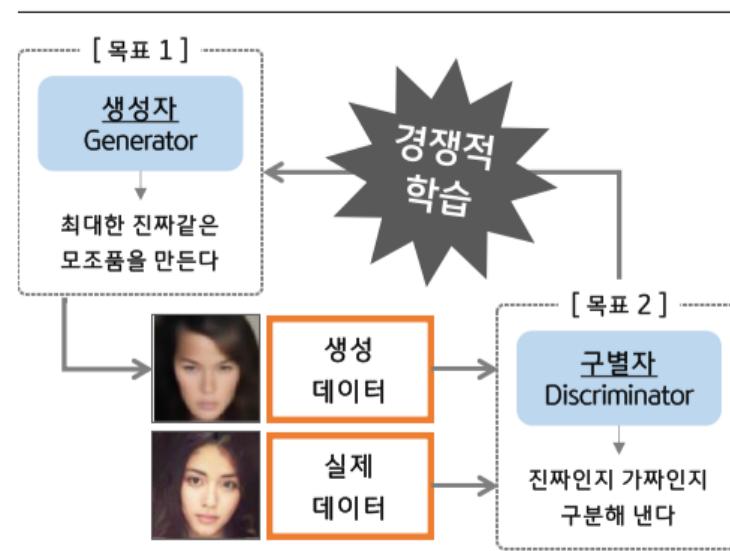


자료: ‘Active Learning in Trading Algorithms(JPMorgan, 2016)’ 및 KB경영연구소

8. 머신러닝 활용

□ 사기 발견

- 두 모델을 경쟁시킴: 생성모델과 구별모델
- 생성모델은 사기와 유사한 데이터를 만들어냄
- 구별모델은 생성모델이 만든 가짜 데이터와 진짜 사기 데이터를 구분
- 창과 방패의 대결: 데이터가 부족할 때 좋은 선택지



자료: Tech M 및 KB경영연구소

8. 머신러닝 활용

□ 신용평가 시스템

- “왜”를 알고싶음
- 특정변수에 약간의 변화를 주어 어떤 변수가 가장 영향력있는 변수인지를 판별
- FICO: 신용평가회사
“Reasons Reporter”를 개발



회색으로 가려진 부분은 해당 이미지가 입력값에서 삭제된 것
얼굴과 눈 부분이 있을 때 개구리로 분류할 확률이 증가(85%)하고
없을 때 거의 0으로 감소, ‘얼굴과 눈’이 개구리 분류의 핵심 요인

자료: Marco Tulio Ribeiro et al.(O'REILLY, 2016)

00. Epilogue

THANK YOU
FOR LISTENING