

사례를 통해 배우는 빅데이터와 머신러닝

김 승 환

swkim4610@inha.ac.kr

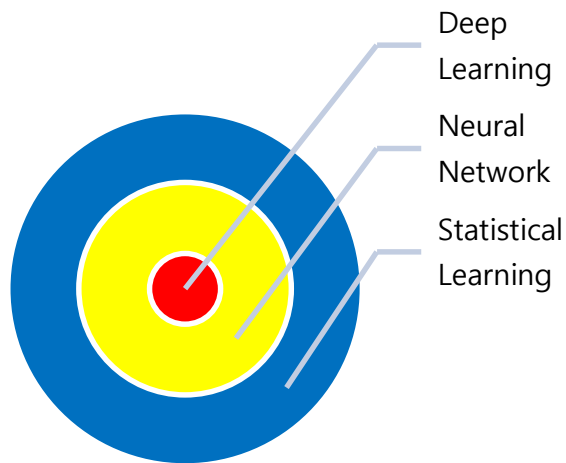
목차

- 0. 빅데이터와 머신러닝
 - 1. 암세포 영상을 이용한 암 진단
 - 2. 스팸메일 차단기
 - 3. 회귀분석과 딥러닝
 - 4. 숫자 인식
 - 5. 음성 인식
 - 6. 강화학습
 - 7. 머신러닝에 필요한 기초지식

0. 빅데이터와 머신러닝

- 2006년 제프리 힌튼은 손글씨 숫자를 인식할 수 있는 딥러닝 모형에 관한 논문을 발표
- 인간보다 더 정확한 결과(98%)를 기록하여 인간보다 컴퓨터가 더 잘할 수 있다는 가능성을 보임
- 이후, 고사양 컴퓨터의 성능과 빅데이터가 결합되어 다른 머신러닝 기법 보다 좋은 결과를 나타내는 여러 사례가 나타남
- 지금 딥러닝은 전세계에서 가장 핫(Hot)한 분야임
- 휴대폰 지문인식, 음성인식, 인공지능 비서에서 자율주행 자동차까지 머신러닝의 분야는 엄청난 발전을 하고 있음
- Machine Learning: 명시적 프로그래밍 없이 컴퓨터 스스로 학습할 수 있는 능력을 갖추게 하는 분야

Machine Learning



0. 빅데이터와 머신러닝

- Supervised Learning:

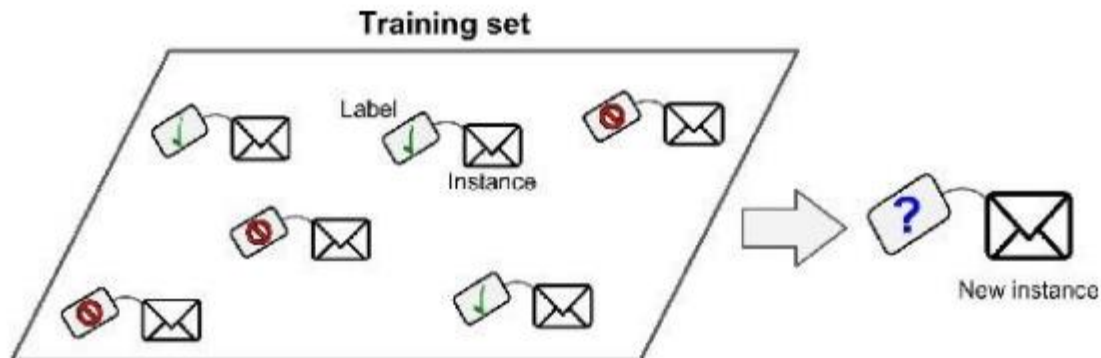


Figure 1-5. A labeled training set for supervised learning (e.g., spam classification)

A typical supervised learning task is *classification*. The spam filter is a good example of this: it is trained with many example emails along with their *class* (spam or ham), and it must learn how to classify new emails.

- k-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests
- Neural networks²

0. 빅데이터와 머신러닝

- Un-supervised Learning:

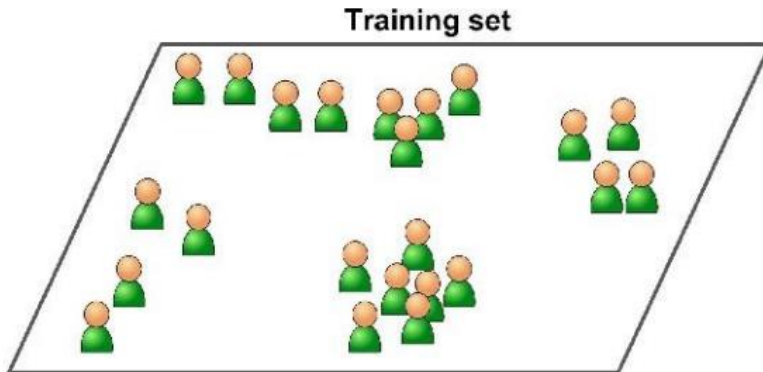


Figure 1-7. An unlabeled training set for unsupervised learning

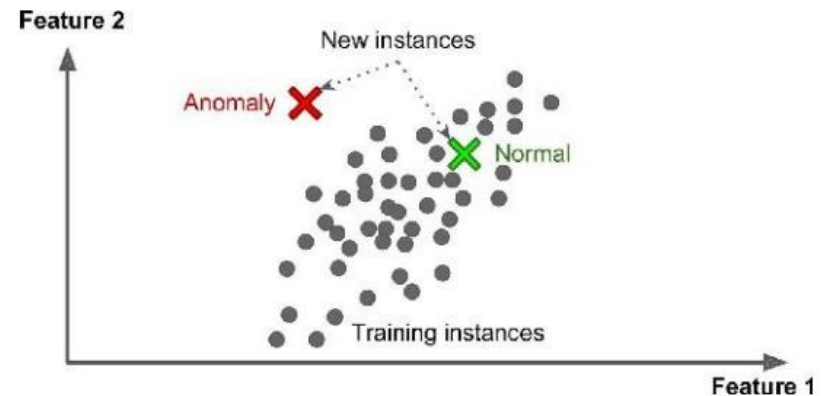


Figure 1-10. Anomaly detection

- Clustering
 - k-Means
- Association rule learning
 - Apriori
- Visualization and dimensionality reduction
 - Principal Component Analysis (PCA)

0. 빅데이터와 머신러닝

- Reinforcement Learning:

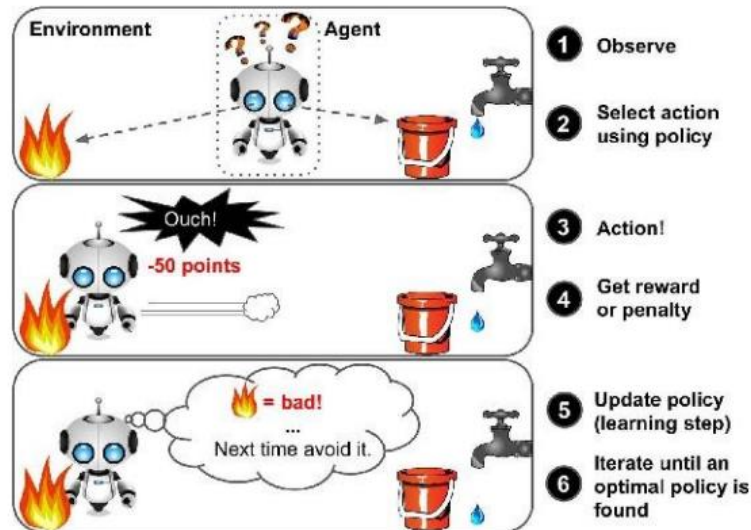
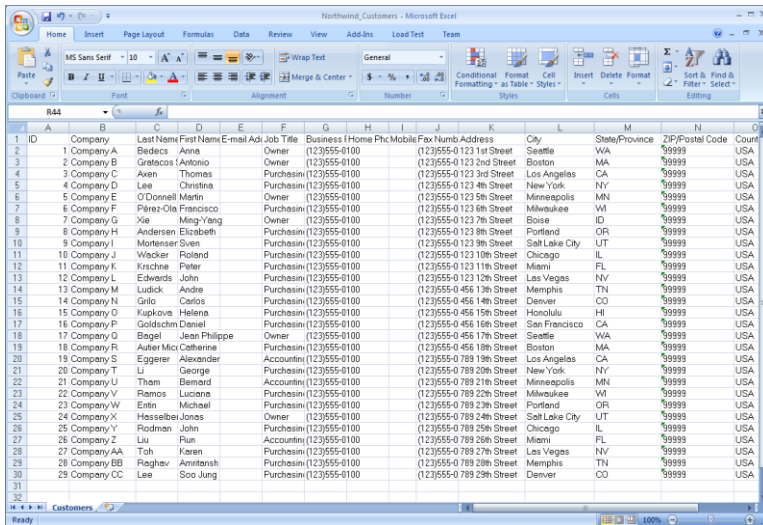


Figure 1-12. Reinforcement Learning

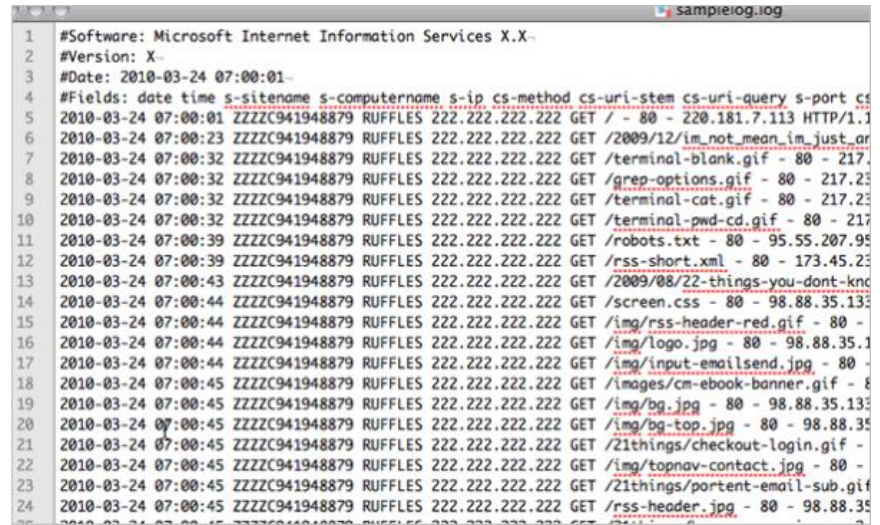
현재 policy에 의해 의사결정 하고 실행한다.
 실행 결과에 따라 벌점을 받을 수 있다.
 벌점을 최소화하는 방향으로 의사결정 policy를 수정한다.
 이 과정을 반복한다.

0. 빅데이터와 머신러닝



ID	Company	Last Name	First Name	E-mail Address	Job Title	Business Home Phone	Mobile Phone Number	Address	City	State/Province	ZIP/Postal Code	Count
1	Company A	Bedeck	Anna	Owner	(12)955-0100	(12)955-0123	1st Street	Seattle	WA	99999	USA	1
2	Company B	Gratcos	Antonio	Owner	(12)955-0100	(12)955-0123	2nd Street	Boston	MA	99999	USA	1
3	Company C	Awen	Thomas	Purchasin	(12)955-0100	(12)955-0123	3rd Street	Los Angeles	CA	99999	USA	1
4	Company D	Lae	Christina	Purchasin	(12)955-0100	(12)955-0123	4th Street	New York	NY	99999	USA	1
5	Company E	O'Donnell	Marlin	Owner	(12)955-0100	(12)955-0123	5th Street	Minneapolis	MN	99999	USA	1
6	Company F	Pérez-Ota	Francisco	Purchasin	(12)955-0100	(12)955-0123	6th Street	Milwaukee	WI	99999	USA	1
7	Company G	Xie	Ming-Yang	Owner	(12)955-0100	(12)955-0123	7th Street	Boise	ID	99999	USA	1
8	Company H	Andersen	Elizabeth	Purchasin	(12)955-0100	(12)955-0123	8th Street	Portland	OR	99999	USA	1
9	Company I	Mortensen	Sven	Purchasin	(12)955-0100	(12)955-0123	9th Street	Salt Lake City	UT	99999	USA	1
10	Company J	Wacker	Roland	Purchasin	(12)955-0100	(12)955-0123	10th Street	Chicago	IL	99999	USA	1
11	Company K	Kirchne	Peter	Purchasin	(12)955-0100	(12)955-0123	11th Street	Miami	FL	99999	USA	1
12	Company L	Edwards	John	Purchasin	(12)955-0100	(12)955-0123	12th Street	Las Vegas	NV	99999	USA	1
13	Company M	Ludick	Andre	Purchasin	(12)955-0100	(12)955-0456	13th Street	Memphis	TN	99999	USA	1
14	Company N	Grilo	Carlos	Purchasin	(12)955-0100	(12)955-0456	14th Street	Denver	CO	99999	USA	1
15	Company O	Kupkova	Helena	Purchasin	(12)955-0100	(12)955-0456	15th Street	Honolulu	HI	99999	USA	1
16	Company P	Goldschm	Daniel	Purchasin	(12)955-0100	(12)955-0456	16th Street	San Francisco	CA	99999	USA	1
17	Company Q	Bagel	Jean Philippe	Owner	(12)955-0100	(12)955-0456	17th Street	Seattle	WA	99999	USA	1
18	Company R	Audier	Mc Catherine	Purchasin	(12)955-0100	(12)955-0456	18th Street	Boston	MA	99999	USA	1
19	Company S	Eggerer	Alexander	Account	(12)955-0100	(12)955-0789	19th Street	Los Angeles	CA	99999	USA	1
20	Company T	Li	George	Purchasin	(12)955-0100	(12)955-0789	20th Street	New York	NY	99999	USA	1
21	Company U	Tham	Bernard	Account	(12)955-0100	(12)955-0789	21th Street	Minneapolis	MN	99999	USA	1
22	Company V	Ramos	Luciana	Purchasin	(12)955-0100	(12)955-0789	22th Street	Milwaukee	WI	99999	USA	1
23	Company W	Enin	Michael	Purchasin	(12)955-0100	(12)955-0789	23th Street	Portland	OR	99999	USA	1
24	Company X	Hasselbe	Jones	Owner	(12)955-0100	(12)955-0789	24th Street	Salt Lake City	UT	99999	USA	1
25	Company Y	Rodman	John	Purchasin	(12)955-0100	(12)955-0789	25th Street	Chicago	IL	99999	USA	1
26	Company Z	Liu	Fun	Account	(12)955-0100	(12)955-0789	26th Street	Miami	FL	99999	USA	1
27	Company AA	Tah	Karen	Purchasin	(12)955-0100	(12)955-0789	27th Street	Las Vegas	NV	99999	USA	1
28	Company BB	Raghav	Amritansh	Purchasin	(12)955-0100	(12)955-0789	28th Street	Memphis	TN	99999	USA	1
29	Company CC	Lee	Soo Jung	Purchasin	(12)955-0100	(12)955-0789	29th Street	Denver	CO	99999	USA	1

정형 Data



```

1 #Software: Microsoft Internet Information Services X.X-
2 #Version: X-
3 #Date: 2010-03-24 07:00:01-
4 #Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-status
5 2010-03-24 07:00:01 ZZZZC941948879 RUFFLES 222.222.222.222 GET / - 80 - 220.181.7.113 HTTP/1.1
6 2010-03-24 07:00:23 ZZZZC941948879 RUFFLES 222.222.222.222 GET /2009/12/im_not_mean_im_just_an
7 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /terminal-blank.gif - 80 - 217.2
8 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /grep-options.gif - 80 - 217.2
9 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /terminal-cat.gif - 80 - 217.2
10 2010-03-24 07:00:32 ZZZZC941948879 RUFFLES 222.222.222.222 GET /terminal-pwd-cd.gif - 80 - 217
11 2010-03-24 07:00:39 ZZZZC941948879 RUFFLES 222.222.222.222 GET /robots.txt - 80 - 95.55.207.95
12 2010-03-24 07:00:39 ZZZZC941948879 RUFFLES 222.222.222.222 GET /rss-short.xml - 80 - 173.45.23
13 2010-03-24 07:00:43 ZZZZC941948879 RUFFLES 222.222.222.222 GET /2009/08/22-things-you-dont-knc
14 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /screen.css - 80 - 98.88.35.133
15 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/rss-header-red.gif - 80 -
16 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/logo.jpg - 80 - 98.88.35.1
17 2010-03-24 07:00:44 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/input-emailsend.jpg - 80 -
18 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /images/cm-ebook-banner.gif - 8
19 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/bg.jpg - 80 - 98.88.35.133
20 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/bg-top.jpg - 80 - 98.88.35
21 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /21things/checkout-login.gif -
22 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /img/topnav-contact.jpg - 80 -
23 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /21things/portent-email-sub.gif
24 2010-03-24 07:00:45 ZZZZC941948879 RUFFLES 222.222.222.222 GET /rss-header.jpg - 80 - 98.88.35

```

Log Data

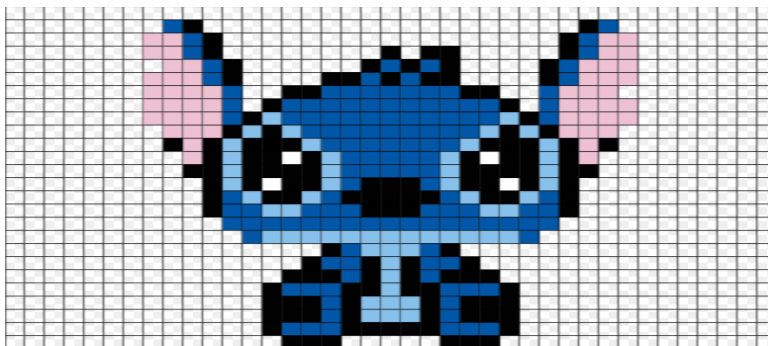
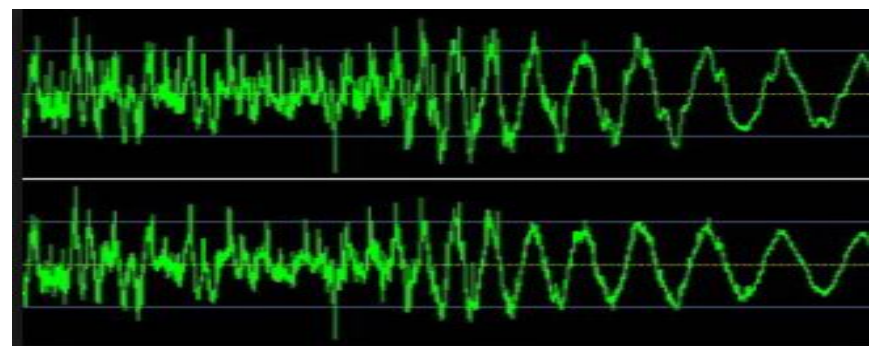


Image Data



Sound Data

0. 빅데이터와 머신러닝

- Static Data / Event Log Data

Static Data는 개체의 속성에 해당하는 데이터로 시간에 따라 바뀌지 않는 데이터를 말함

예: 성별, 연령, 지역, 제조사, 생산일 등

Event Log Data는 개체의 상태에 해당하는 데이터로 시간에 따라 바뀌는 데이터를 말함

예: 현 위치, 조회 키워드, 클릭 페이지 등

일반적으로 행은 관측단위, 열은 변수로 지정되는 정형 데이터의 구조에서는

Event Log 데이터를 처리하기 어려움

하지만, 현재 Event Log를 이용한 데이터 처리 수요가 증가하고 있음

예: 구글, 페이스북 광고 추천, Facebook 자살 예측, 센서 데이터에 의한 Alert 시스템 등

0. 빅데이터와 머신러닝

간단한 문제: 인하대학교 내에 키 185cm 이상인 사람의 비율은?

솔루션 1: 30명의 키를 검사하여 185cm 이상인 사람의 비율로 추정

솔루션 2: 185cm 이상인 사람이 나올 때까지 검사하여 비율은 $1/n$ 으로 추정

솔루션 3: 30명의 키를 측정하여 평균과 표준편차를 구해 정규분포 가정 하에 계산함

솔루션 4: 10,000명의 키를 검사하여 185cm 이상인 사람의 비율로 추정

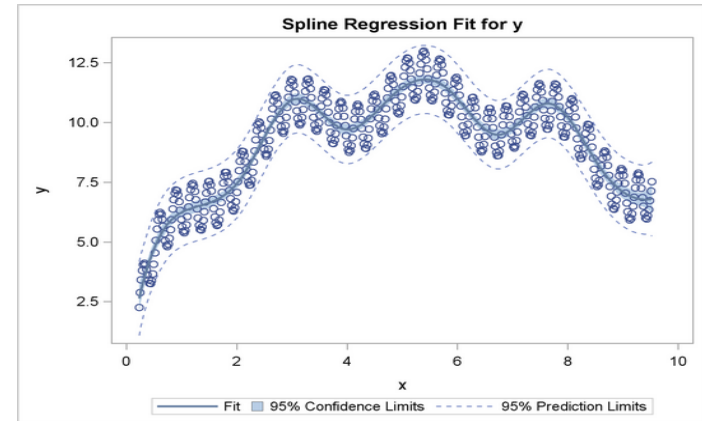
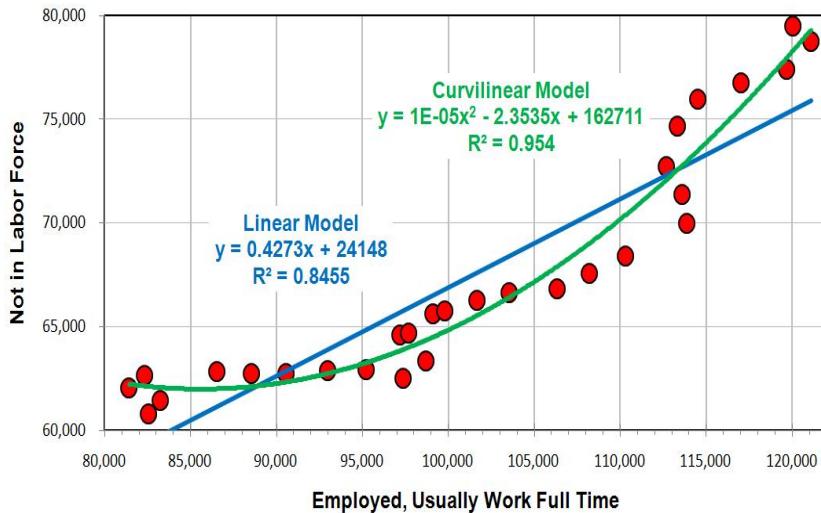
가장 정확한
방법은?



가장 효율적인
방법은?

0. 빅데이터와 머신러닝

Linear Regression and Non-Linear Regression Model

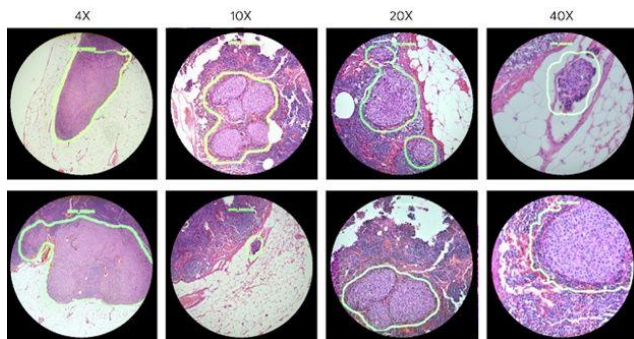


스몰 데이터에서 고차항의 모형을 할 수는 있으나 Over-fitting 위험이 커진다.

빅데이터에서는 복잡한 모형을 사용해도 Over-fitting 위험이 덜하다.

1. 암세포 영상을 이용한 암 진단(k-NN 알고리즘)

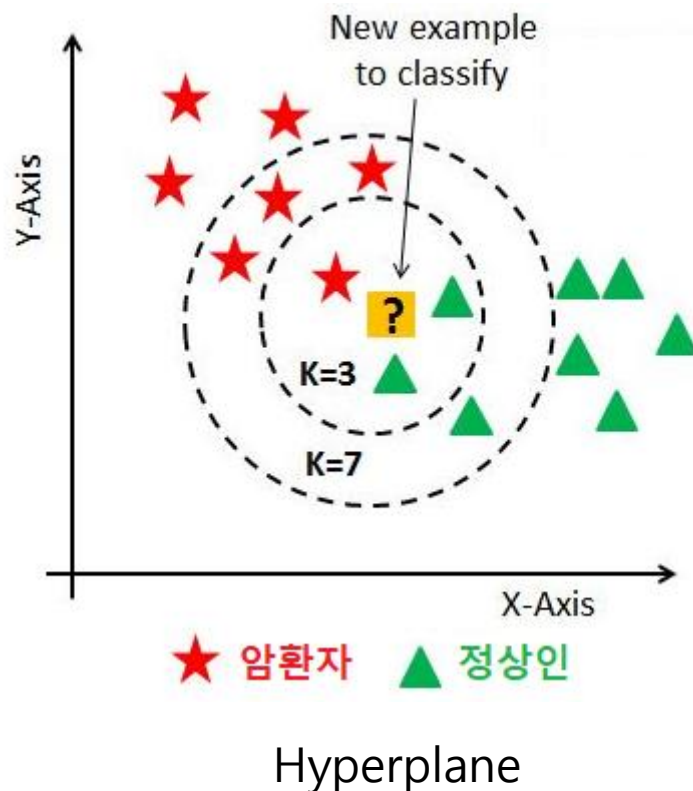
- 암으로 판정된 영상과 정상으로 판정된 영상 자료를 빅데이터로 인정될 만큼 많이 확보한다.
- Computer Vision 기술을 이용하여 영상자료로부터 수치 특징정보를 구한다.
- 우측과 같이 암 존재 여부와 영상에서 구한 수치특징 정보를 Hyperplane 공간에 plotting한다.
- 새로운 환자의 수치특징과 가까운 k 명의 결과를 이용해 암여부를 판정한다.



id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	points_mean	symmetry_mean
1	8713940 B	12.320	12.39	78.85	464.1	0.10280	0.06981	0.039870	0.037000	0.1959
2	8910251	10.600	18.95	69.28	346.4	0.09688	0.11470	0.063870	0.026420	0.1922
3	9055320 B	11.040	16.83	70.92	373.2	0.10770	0.07804	0.030460	0.024800	0.1714
4	868871 B	11.280	13.39	73.00	384.8	0.11640	0.11360	0.046350	0.047960	0.1771
5	9012568 B	15.190	13.21	97.65	711.8	0.07963	0.06934	0.033930	0.026570	0.1721
6	906539 B	11.570	19.04	74.20	409.7	0.08546	0.07722	0.054850	0.014280	0.2031
7	925291 B	5.10	23.93	74.52	403.5	0.09261	0.10210	0.111200	0.041050	0.1388
8	87880 M	23.75	91.56	597.8	0.13230	0.155800	0.091760	0.2251		
9	862989 B	19.29	67.41	336.1	0.09989	0.0778	0.029950	0.012010	0.2217	
10	89827 B	71.49	373.9	0.10330	0.053970	0.033410	0.1776			
11	8	320.0	0.10850	0.023310	0.1970					
12	91131	536.9	0.1562							
13	857810 B	82.61	0.1819							
14	911805 M	18.590	25.00	127.70	0.1663					
15	925277 B	14.590	22.68	96.39	0.1454					
16	867387 B	15.710	13.93	102.00	0.1816					
17	89511502 B	12.670	17.30	81.25	489.9	0.1707				
18	89263202 M	20.090	23.86	134.70	1247.0	0.2249				
19	866714 B	12.190	13.29	79.08	455.8	0.10660	0.09509	0.028550	0.028820	0.1880
20	874373 B	11.710	17.19	74.68	420.3	0.09774	0.06141	0.038090	0.032390	0.1516

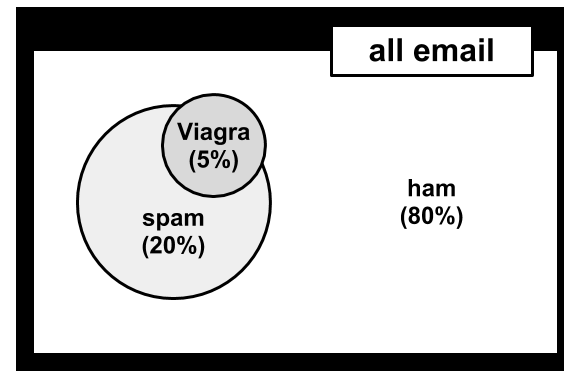
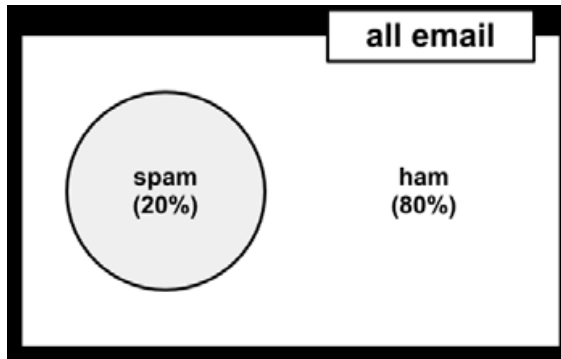
암 존재 여부

조직의 반지름,
면적, 평평도,
굴곡, 대칭성 등



2. 베이즈 정리를 이용한 스팸 메일 차단기

- 우리 메일 중에 햄 메일이 더 많지만, 메일에 ‘viagra’ 라는 단어가 들어가 있다면 스팸 가능성 ↑
- 어떤 사건이 일어난다는 전제 하에 확률을 구하는 것을 조건부 확률이라고 한다.



Bayes Law
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

$$P(\text{spam} | \text{Viagra}) = \frac{P(\text{Viagra} | \text{spam}) P(\text{spam})}{P(\text{Viagra})}$$

Diagram labels for the equation above:
 - **posterior probability** points to $P(\text{spam} | \text{Viagra})$
 - **likelihood** points to $P(\text{Viagra} | \text{spam})$
 - **prior probability** points to $P(\text{spam})$
 - **marginal likelihood** points to $P(\text{Viagra})$

2. 베이즈 정리를 이용한 스팸 메일 차단기

- 스팸여부를 구분할 수 있는 여러 단어를 이용해 조건부 확률을 구해보자.
예를 들어, $W_1 = \text{'Viagra'}$, $W_2 = \text{'money'}$, $W_3 = \text{'Groceries'}$, $W_4 = \text{'Unsubscribe'}$ 를 이용해 스팸일 확률을 계산할 수 있다.
- 문제는 이러한 단어가 많을수록 정확한데 많을수록 계산이 힘들어진다.
모든 단어가 다 들어 있는 문서가 희소하기 때문에 확률계산을 위해서는 아주 많은 메일을 학습해야 함
- 이 문제를 해결하기 위해 가정을 단어끼리 메일에 나올 확률이 독립이라는 나이브한 가정을 해보자.
- 독립을 가정하기 때문에 정확도 손실이 발생하지만, 독립을 가정하지 않아도 정확한 계산이 힘들기 때문에 독립을 가정하고 계산한다.

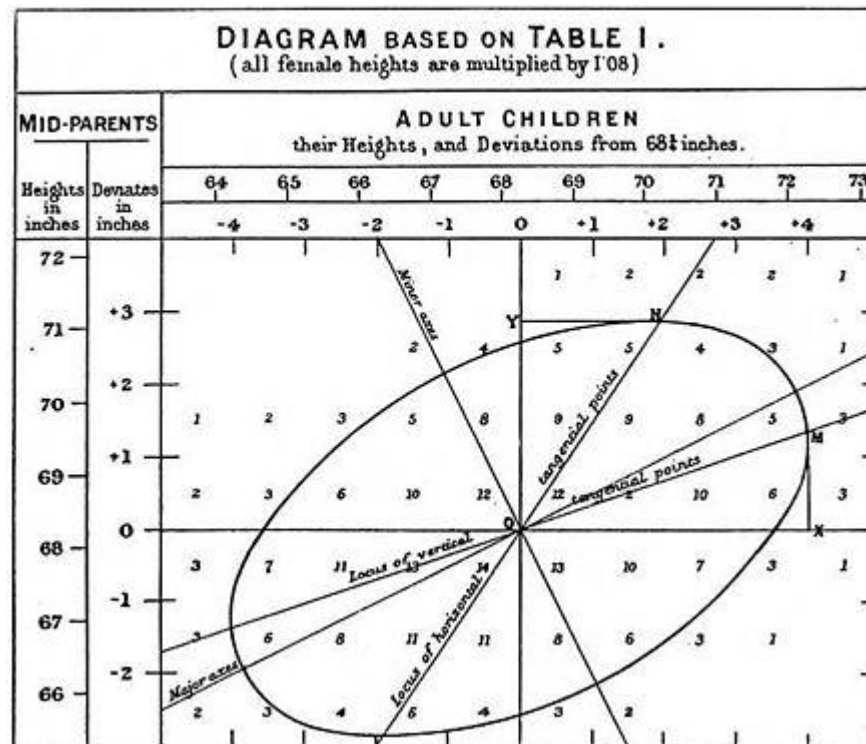
Using Bayes' theorem, we can define the problem as shown in the following formula, which captures the probability that a message is spam, given that Viagra = Yes, Money = No, Groceries = No, and Unsubscribe = Yes:

$$P(\text{Spam} | W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4 | \text{spam}) P(\text{spam})}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

$$P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4 | \text{spam}) \xrightarrow{\text{naive}} P(W_1 | \text{spam}) P(\neg W_2 | \text{spam}) P(\neg W_3 | \text{spam}) P(W_4 | \text{spam})$$

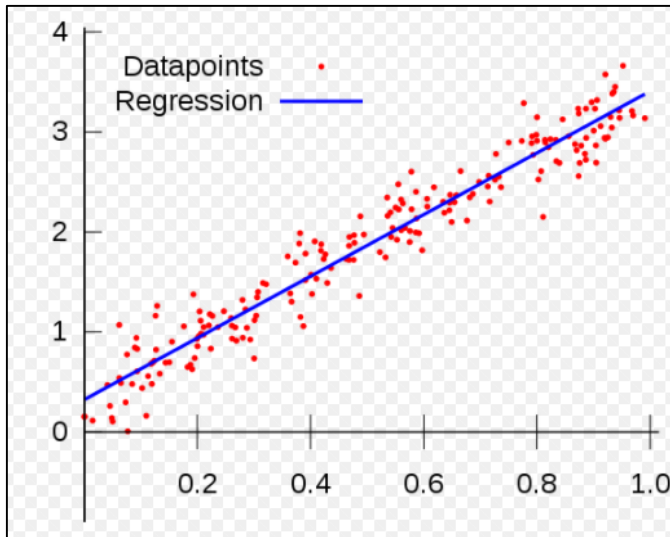
3. 회귀분석과 답러닝

[Francis Galton's](#) 1875 illustration of the correlation between the heights of adults and their parents. The observation that adult children's heights tended to deviate less from the mean height than their parents suggested the concept of "[regression toward the mean](#)", giving regression its name.

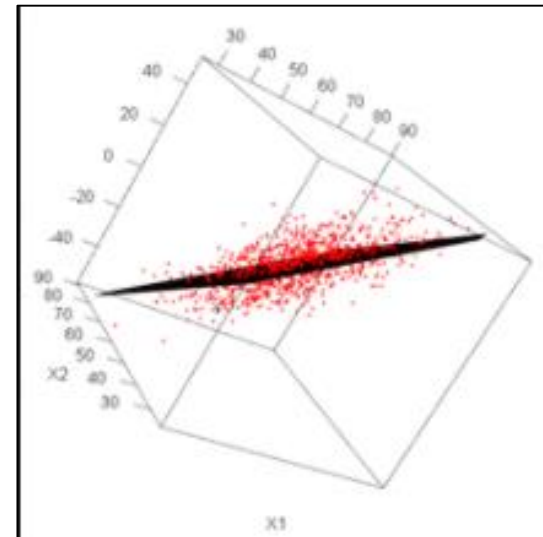


3. 회귀분석과 딥러닝

회귀분석은 관측 값을 가장 잘 지나가는 직선 혹은 곡선의 방정식을 구하는 방법론



$$Y = \alpha + \beta x + \epsilon$$



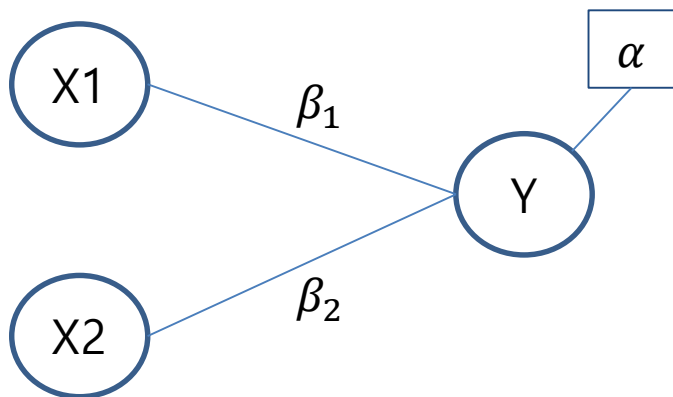
$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

직선으로 예측했을 경우, 오차 ϵ 이 존재, 오차가 작을 수록 좋은 모형임

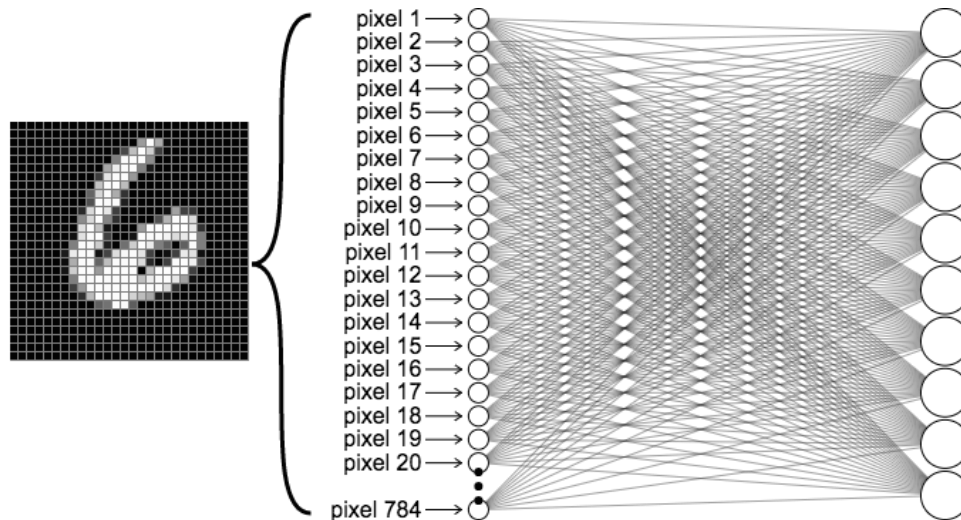
3. 회귀분석과 딥러닝

회귀분석의 정확도를 높이기 위해 아주 복잡한 형태의 비선형 회귀모형을 만든다.
 모형이 복잡해지면 학습자료에서는 잘 맞지만, 실제상황에서는 부정확해지는 문제 발생
 → 빅데이터 시대가 되면서 이러한 문제가 어느 정도 해결됨

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$



$$Y = f(p_1, p_2, \dots, p_{784})$$



딥러닝은 아주 복잡한 비선형 회귀분석 모형임

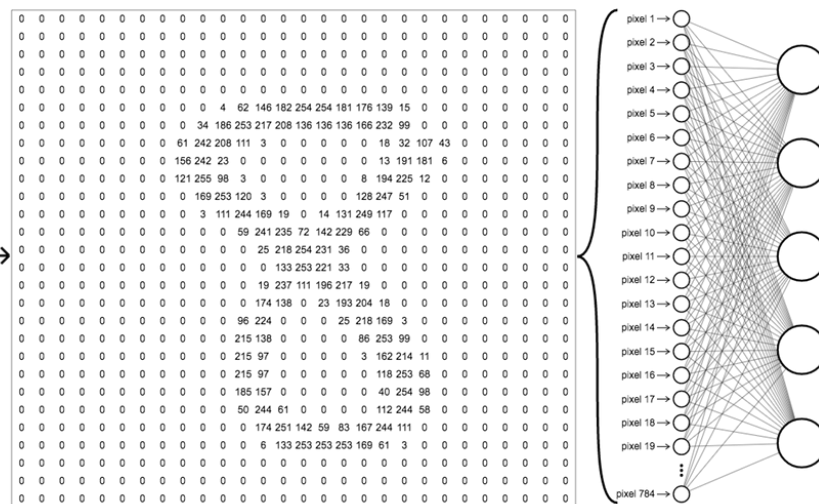
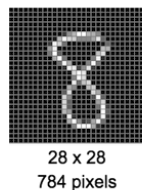
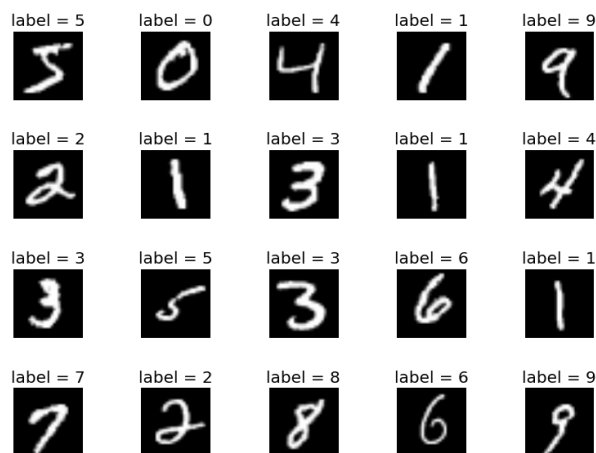
3. 회귀분석과 딥러닝

농구에서 일반적인 3점 슈트가 있고 공중에서 패스를 받아 슈트를 시도하는 앨리웁 슈트가 있다. 회귀분석모형을 3점 슈트라고 하면 앨리웁 슈트는 히든 레이어가 있는 딥러닝 모형이라고 할 수 있다.



4. 딥러닝으로 숫자 인식하기

1. 우편번호, 자동차 번호판 인식 등 컴퓨터를 이용해 숫자를 인식하는 것은 실생활에서 많이 볼 수 있다.
2. 숫자/문자 인식은 아래와 같이 픽셀 값을 독립 변수로 보고 label을 종속변수로 보는 회귀모형이다.
3. 최근에 필터를 사용해 이미지를 단순화 시켜 정확도를 증가시키는 획기적인 알고리즘이 개발되어 컴퓨터의 숫자/문자 인식 정확도가 사람보다 더 정확하다.(Convolutional Neural Network)
4. CNN 알고리즘의 개발로 전세계의 딥러닝 열풍이 시작되었다.
5. 이 원리를 이용해 수많은 지문인식, 얼굴인식 등 수 많은 이미지 처리 서비스가 등장하였다.



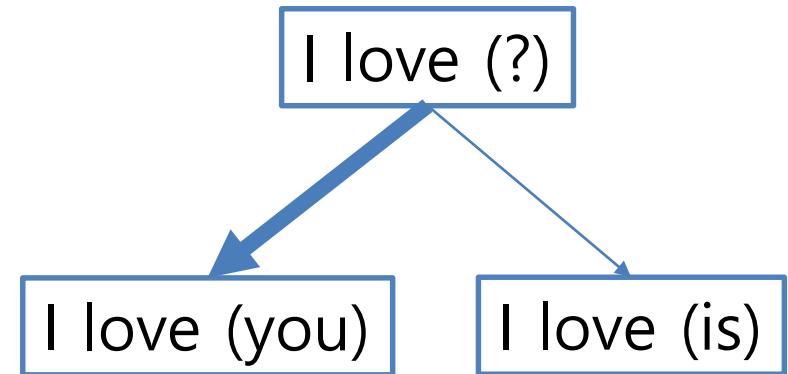
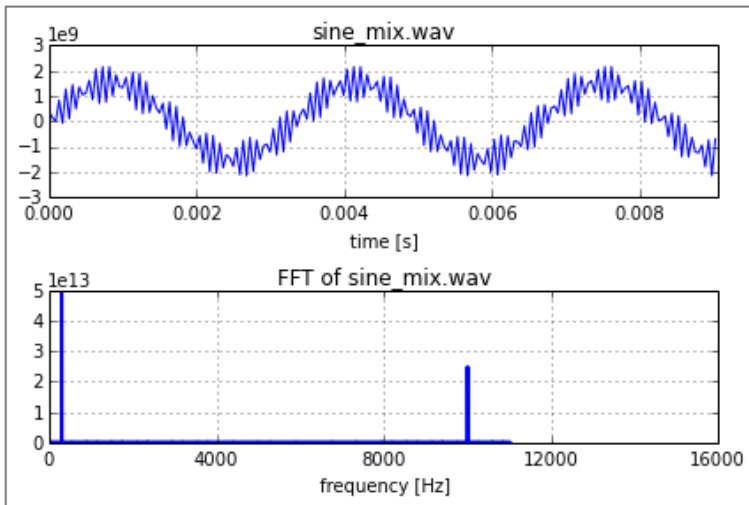
5. 음성 인식

1. 음성 / 소리는 마이크 떨림 위치를 시간에 따라 저장한 디지털 정보(wave file format)를 이용한다.
2. 소리는 연속적이므로 이를 잘게 나누어 샘플링하고 소리의 특징을 추출하여 학습한다.
3. 예를 들어, 음성신호 Dog, Cat을 독립변수로 보고 종속변수를 영어 단어 Dog, Cat으로 추정하는 함수를 만드는 것이다.



5. 음성 인식

4. 음성인식의 정확도를 높이기 위해 시간순서인 웨이브 포맷을 후리에 변환을 통해 주파수별 파워로 변환한 다음 주파수별 파워량을 특징으로 보고 이를 딥러닝 모형으로 학습하여 인식한다.
5. 음성인식의 경우, 레이블이 너무 많아 정확도를 높이기 위해서 다른 많은 개선이 필요하다.
6. 문장에서 단어들 간의 연관성을 이용하여 개선하는 방법이 사용되고 있다.
7. AI 비서, 콜센터 자동화 등에 이용한다.



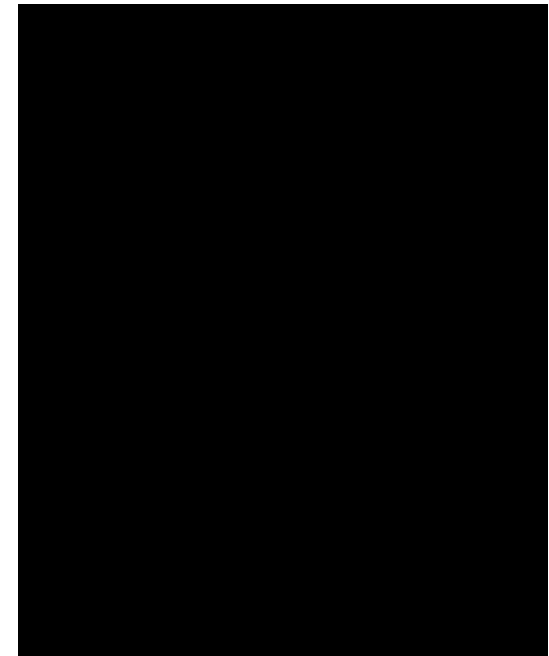
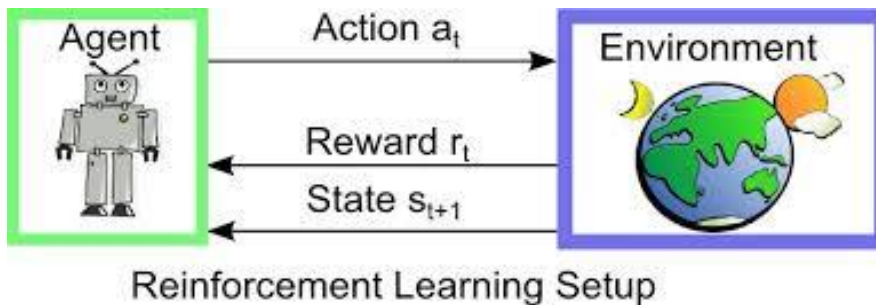
6. 강화 학습

1. 알파고는 이세돌과 대전할 때, 먼저 프로기사의 기보 3,000만건으로 지도학습을 수행하고 셀프 대국으로 강화학습을 하여 인공지능을 향상시켰다.
2. 이후, 구글은 알파제로를 발표했는데 이는 100% 강화학습으로 만들어진 알고리즘이다.



6. 강화 학습

1. 강화 학습은 Agent가 임의의 Action을 통해 보상을 받는 환경을 만들고 이 환경에서 받는 보상이 최대화되도록 자신의 Action을 수정하는 알고리즘이다.
2. 어떤 Action이 보상을 최대화하는지 알기 위해서 모든 가능한 Action을 해봐야 하는데 너무 많다.
3. 이를 줄이기 위해 Action과 Reward의 관계를 딥러닝으로 학습해 시행착오를 줄이는 것이 핵심이다.



7. 머신러닝에 필요한 기초지식

수학, 통계학, IT, Biz. 이해

$$\begin{aligned}
 \text{Cost}(w) &= \sum_{i=1}^n \sum_{j=1}^c (-Y_{ij} \cdot \log H(X)_{ij}) = \sum_{i=1}^n (-Y_{i1} \cdot \log H(X)_{i1} - Y_{i2} \cdot \log H(X)_{i2}) \\
 &= \sum_{i=1}^n (-Y_{i1} \cdot \log H(X)_{i1} - (1 - Y_{i1}) \cdot \log(1 - H(X)_{i1}))
 \end{aligned}$$

```

sess = tf.Session()
sess.run(tf.global_variables_initializer())
print('Learning started. It takes sometime.')

for epoch in range(training_epochs):
    avg_cost = 0
    total_batch = int(mnist.train.num_examples / batch_size)

    for i in range(total_batch):
        batch_xs, batch_ys = mnist.train.next_batch(batch_size)
        feed_dict = {X: batch_xs, Y: batch_ys}
        c, _ = sess.run([cost, optimizer], feed_dict=feed_dict)
        avg_cost += c / total_batch
    print('Epoch:', '%04d' % (epoch + 1), 'cost =', '{:.9f}'.format(avg_cost))
  
```



끝?

끝? 이 아닌 이제 시작이다.
이 내용을 토대로 수많은 실질 문제를 풀어보고,
실전에 적용할 수 있는 칼을 만들어야겠다.

수고 하셨습니다.~~~
 짹 ~ 짹 ~ 짹 ~~~