사례를 통해 배우는 빅데이터와 머신러닝

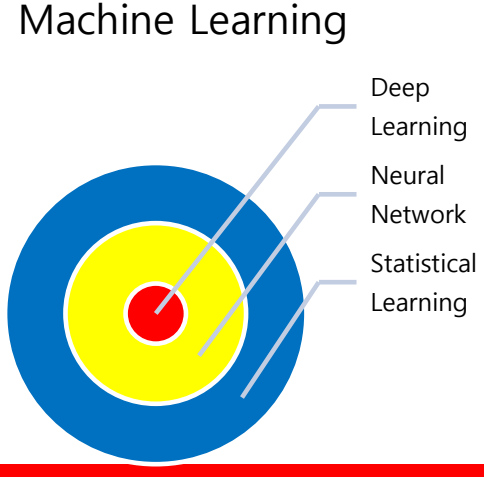
김 승 환

swkim4610@inha.ac.kr

목차

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



0. 빅데이터와 머신러닝

- Supervised Learning: ← 정답을 제공함! ⇒ 여러 예제들에 대한 답을 제공한 후, 새로운 Instance를 판별하는 것 → 새로운 input data.

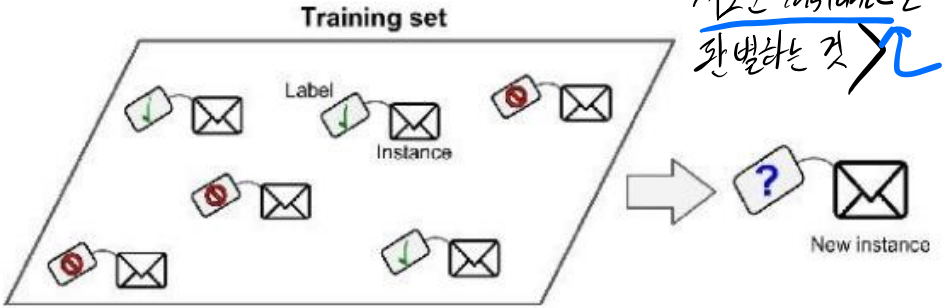


Figure 1-5. A labeled training set for supervised learning (e.g., spam classification)

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 networks[2]

0. 빅데이터와 머신러닝

- Un-supervised Learning: ← 정답을 제공하지 않음. ⇒ 〈정답을 미리 제공하지 않고, 미리 제시된 데이터들의 특성들을 기준으로 집단을 나눈다. 이렇게 생성된 집단의 특성을 기준으로 새로운 instance의 특성과 집단의 특성을 비교하여, 해당 instance를 어느 집단에 넣어야 할지 판단하는 것.〉

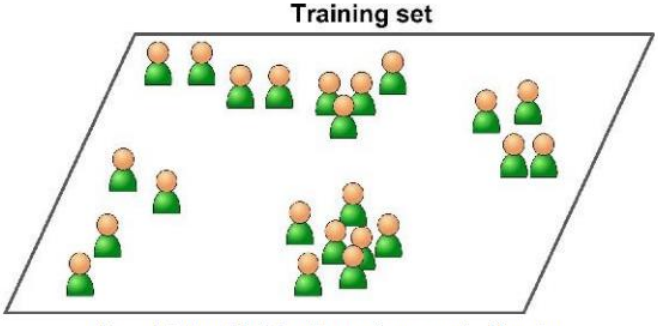


Figure 1-7. An unlabeled training set for unsupervised learning

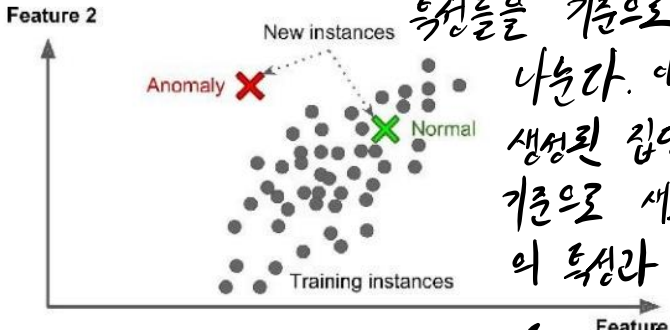


Figure 1-10. Anomaly detection

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

0. 빅데이터와 머신러닝

- Reinforcement Learning: (강화 학습) → 여러 번의 시행착오를 실시함.

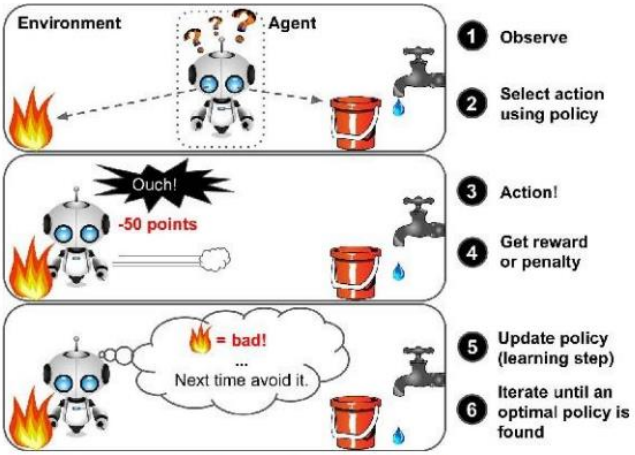


Figure 1-12. Reinforcement Learning

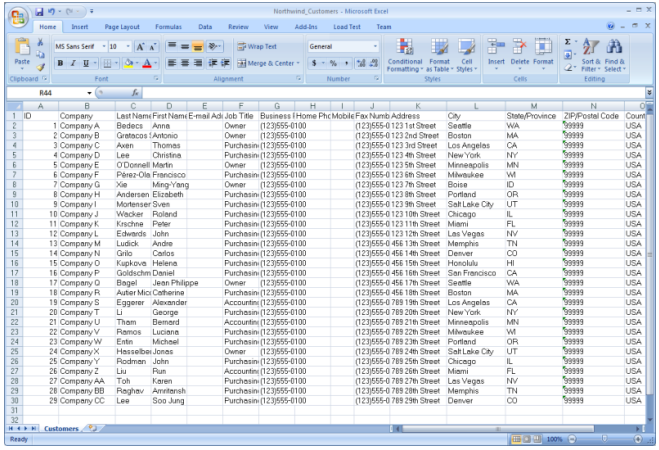
현재 policy에 의해 의사결정 하고 실행한다.
실행 결과에 따라 벌점을 받을 수 있다.
벌점을 최소화하는 방향으로 의사결정 policy를 수정한다.
이 과정을 반복한다.

(`열'이 존재하는 form)

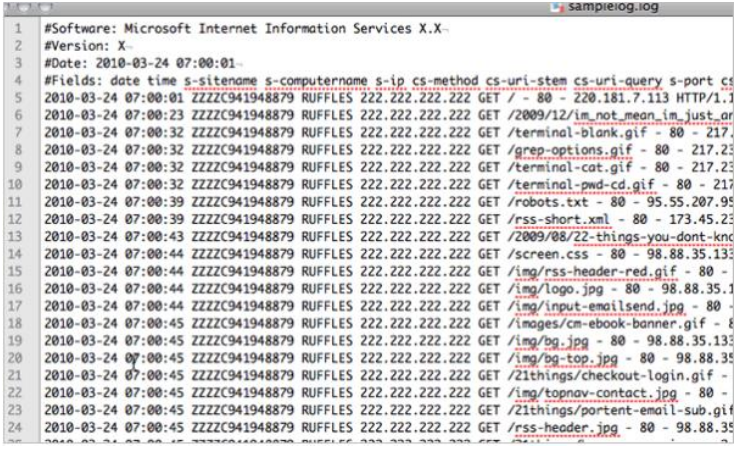
0. 빅데이터와 머신러닝

정형화된 form이 들어있는 data

정형화된 form이 들어갈 수 없는 data.



정형 Data



Log Data

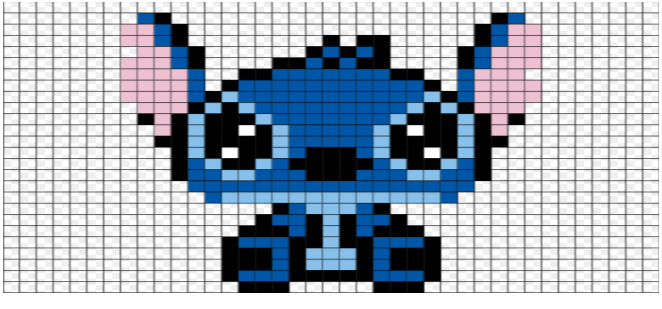
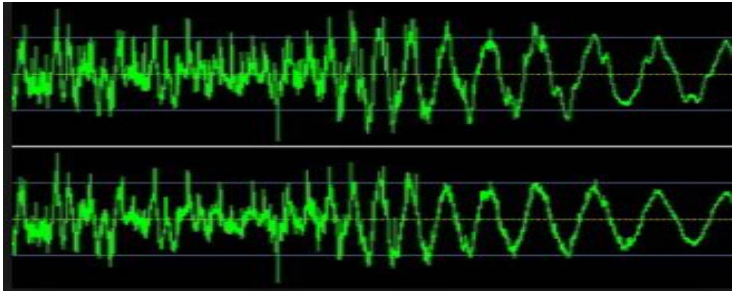


Image Data



Sound Data

0. 빅데이터와 머신러닝

- Static Data / Event Log Data

정적 데이터 (=속성 데이터)

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

예: 성별, 연령, 지역, 제조사, 생산일 등

Event Log Data는 개체의 상태(=행동)에 해당하는 데이터로 시간에 따라 바뀌는 데이터를 말함

예: 현 위치, 조회 키워드, 클릭 페이지 등

행동 데이터

'Deep learning'은 event log data를 예측할 때 활용된다!

일반적으로 행은 관측단위, 열은 변수로 지정되는 정형 데이터의 구조에서는

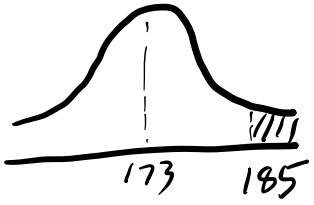
Event Log 데이터를 처리하기 어려움

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

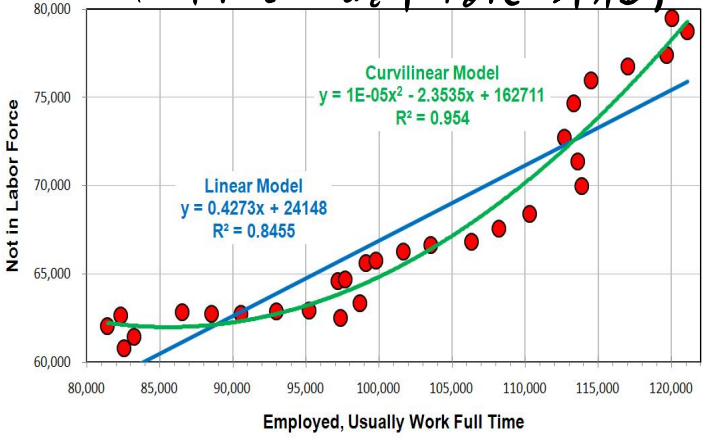
가장 정확한 방법은?

가장 효율적인 방법은?

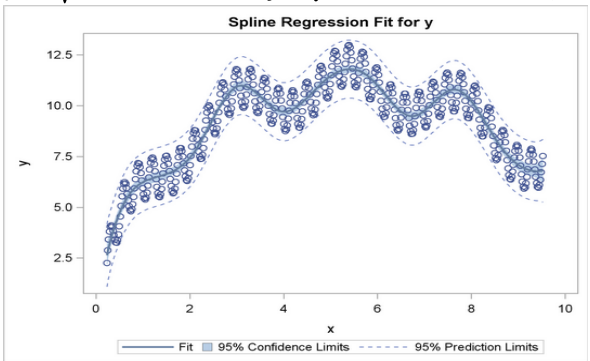
0. 빅데이터와 머신러닝

Linear Regression and Non-Linear Regression Model

↳ 데이터가 없는 상황에 사용하는 회귀분석



↳ 데이터가 많은 상황에 사용할 수있는 회귀분석.

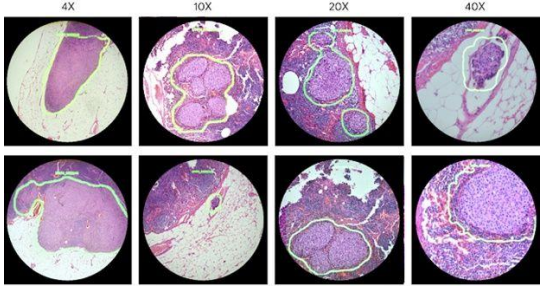


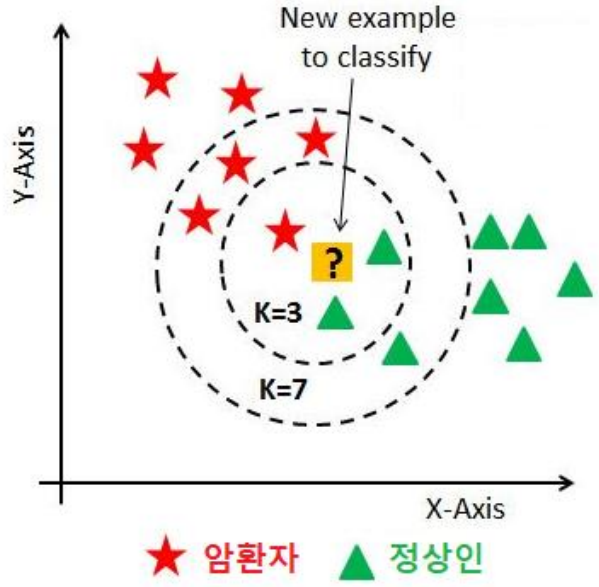
스몰 데이터에서 고차항의 모형을 할 수는 있으나 Over-fitting 위험이 커진다.

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.

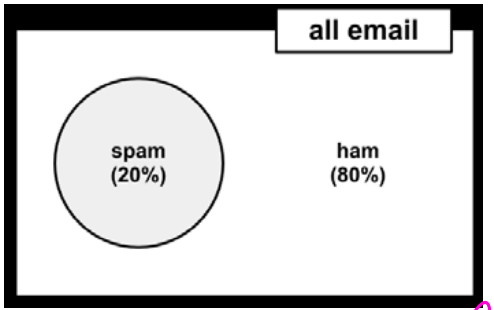


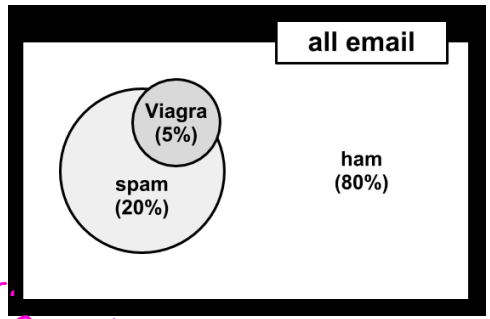


Hyperplane

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.





① 이걸 알면 ② 이걸 알 수 있다.

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

likelihood, prior probability

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

posterior probability, marginal likelihood

2. 베이즈 정리를 이용한 스팸 메일 차단기

- 스팸여부를 구분할 수 있는 여러 단어를 이용해 조건부 확률을 구해보자.
  예를 들어, W1= 'Viagra', W2= 'money', W3= 'Groceries', W4= '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:

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

'없다'는 개념

'독립'

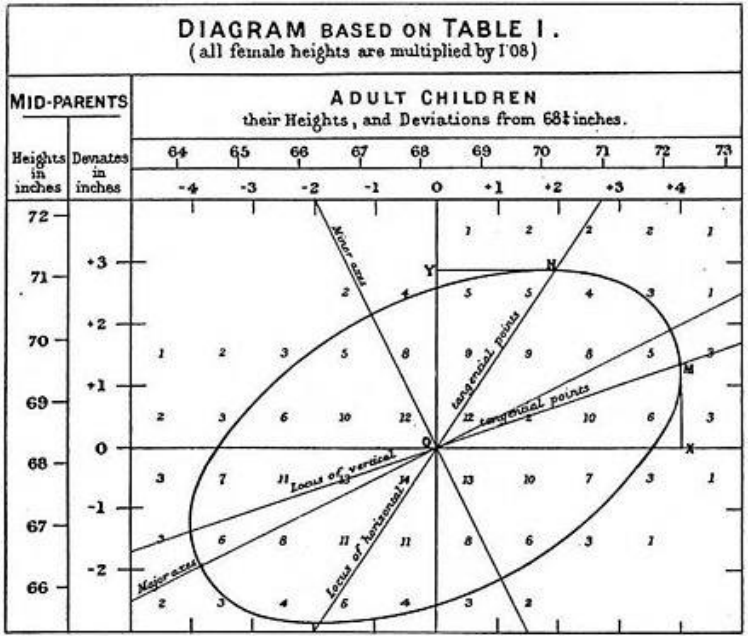
$$P(W_1 \cap !W_2 \cap !W_3 \cap W_4 \mid spam) \xrightarrow{naive\ (순진한,\ 바보인)} P(W_1 \mid spam)P(!W_2 \mid spam)P(!W_3 \mid spam)P(W_4 \mid spam)$$

↑ 해당 사건들을 독립이라고 봤을 때 = 사실, 해당 사건들을 단순히 독립으로 보면 안된다.

그런데, 이 방법으로 계산하면 ... 정확하다.

3. 회귀분석과 딥러닝

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.

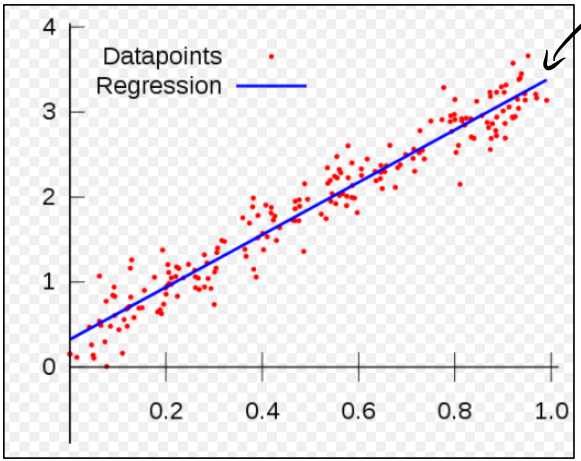


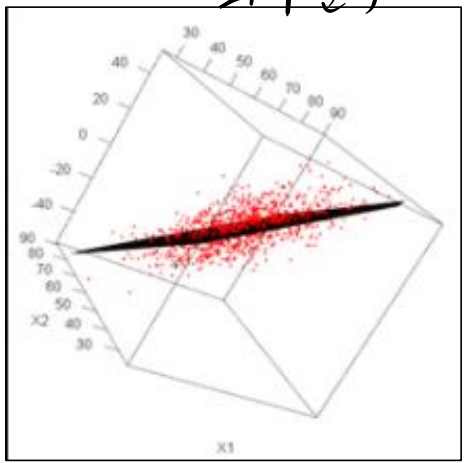
회귀분석의 시초 (유전학)

3. 회귀분석과 딥러닝

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

<해당 방정식을 구하는 것 자체가 '회귀 분석'이다.>





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

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

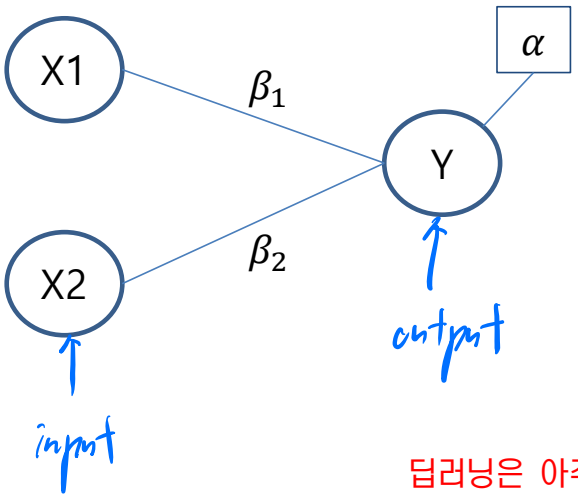
직선으로 예측했을 경우, 오차 ε 이 존재, 오차가 작을 수록 좋은 모형임

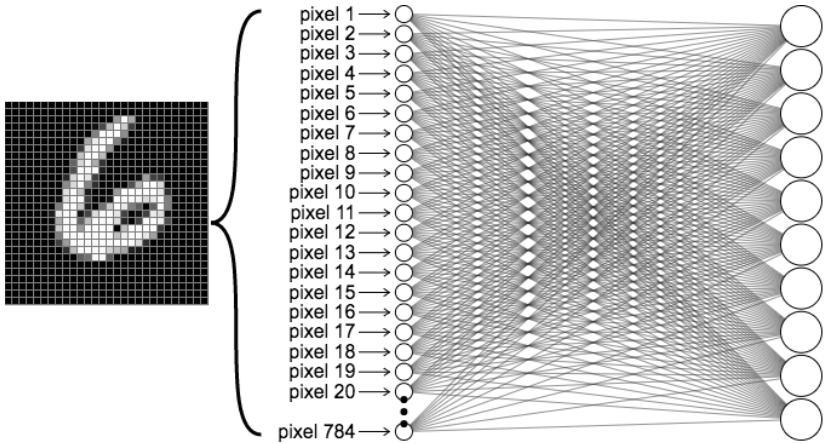
3. 회귀분석과 딥러닝

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

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

$$Y = f(p1, p2, \ldots, p784)$$

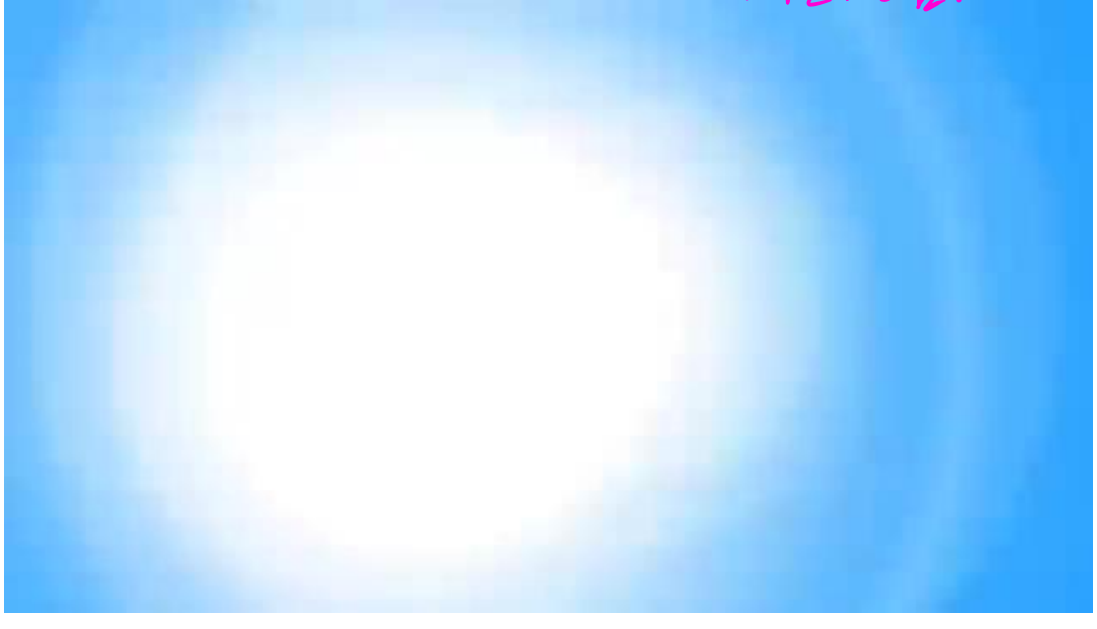




딥러닝은 아주 복잡한 비선형 회귀분석 모형임
⇒ 딥러닝은 결국 '회귀분석'이다!!

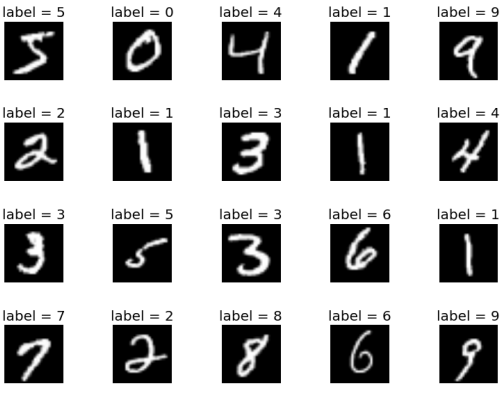
3. 회귀분석과 딥러닝

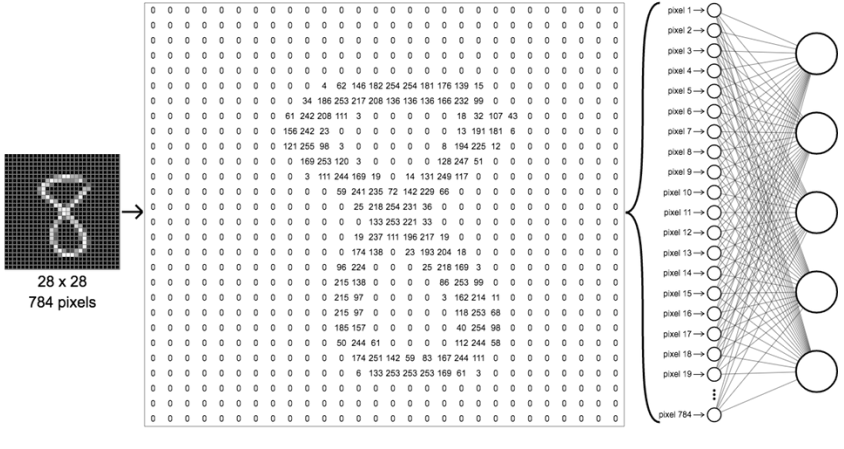
농구에서 일반적인 3점 슛이 있고 공중에서 패스를 받아 슛을 시도하는 앨리웁 슛이 있다. 회귀분석모형을 3점 슛이라고 하면 앨리웁 슛은 히든 레이어가 있는 딥러닝 모형이라고 할 수 있다.



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.



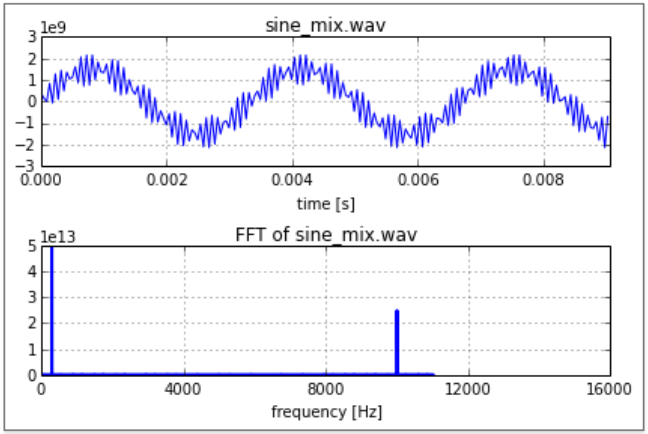


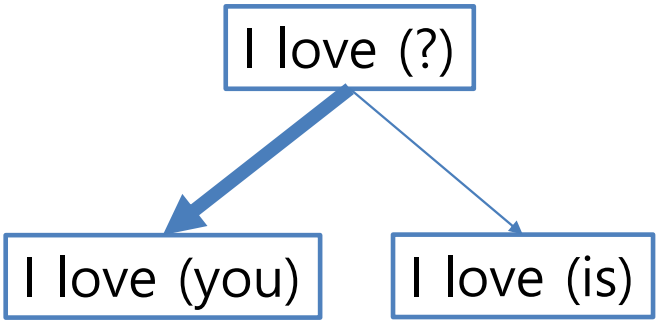
5. 음성 인식

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.



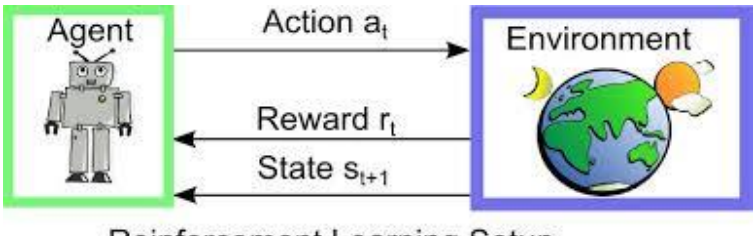


6. 강화 학습

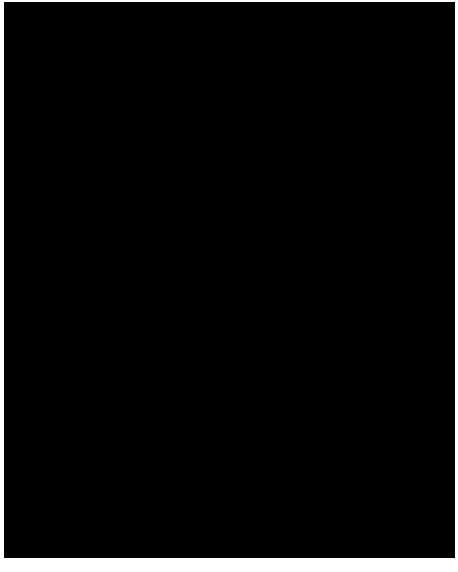
1. 알파고는 이세돌과 대전할 때, 먼저 프로기사의 기보 3,000만건으로 지도학습을 수행하고 셀프 대국으로 강화학습을 하여 인공지능을 향상시켰다.
2. 이후, 구글은 알파제로를 발표했는데 이는 100% 강화학습으로 만들어진 알고리즘이다.

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.



Reinforcement Learning Setup



7. 머신러닝에 필요한 기초지식

수학, 통계학, IT, Biz. 이해

$$Cost(w) = \sum_{i=1}^{n}\sum_{j=1}^{c}\left(-Y_{ij}\cdot logH(X)_{ij}\right) = \sum_{i=1}^{n}\left(-Y_{i1}\cdot logH(X)_{i1} - Y_{i2}\cdot logH(X)_{i2}\right)$$

$$= \sum_{i=1}^{n}\left(-Y_{i1}\cdot logH(X)_{i1} - (1-Y_{i1})\cdot \log(1-H(X)_{i1})\right)$$

```
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.~~~

짝 ~ 짝 ~ 짝 ~~~