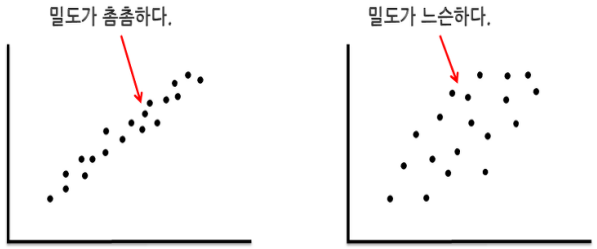
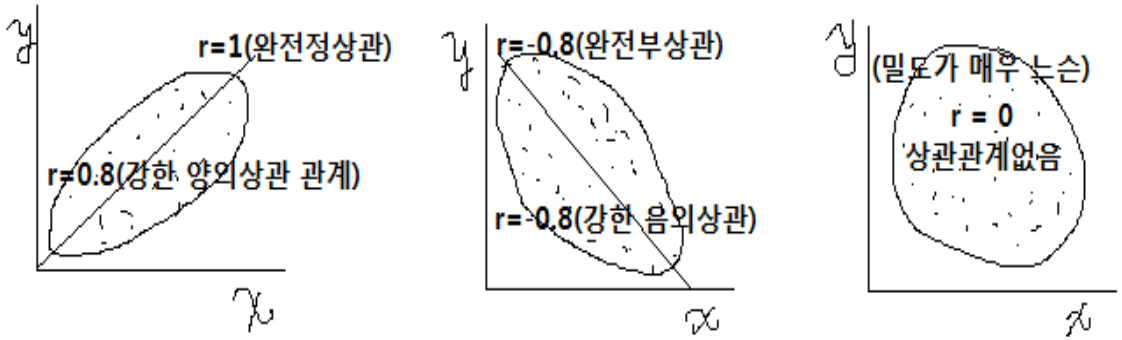
상관관계 분석 (Correlation Analysis)

- 회귀분석에서 변수들 간의 인과관계를 분석하기 전에 각 변수들 간에 관련성을 분석하는 선행자료로 이용된다.
- 데이터 간의 밀도는 상관계수의 수치를 사용하여 관계의 정도를 파악할 수 있다.
- 변수 간 관련성 분석, 관계의 친밀함을 수치로 표현할 수 있다.
  예) 광고비와 매출액 사이의 관련성 분석, 광고량과 브랜드 인지도의 관련성 분석



*** 상관계수 r과 상관관계 정도**

- 상관계수는 밀도를 숫자로 표현한다. 밀도를 가지고 상관관계를 정확하게 표현하기 힘들다.
  그래서 숫자화 해야 한다. 이 것을 정도에 따라 구분한 것 중 하나가 피어슨 상관계수다. 이는 두 변수 간의 관련성을 알기 위해 이용된다.
- 상관계수 r은 -1 ~ 1 사이의 값을 갖는다. (1 : 완전상관(밀도 촘촘), 0 : 상관관계 없다)



x가 커지면 y도 커진다(정비례)
x가 커질수록 y는 작아진다(반비례)

상관관계 분석

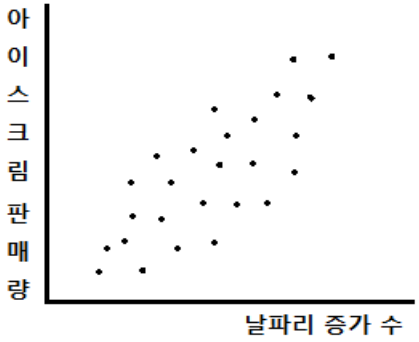
상관분석은 변수 간에 어떠한 관계가 있는지 상관관계를 파악할 수는 있지만, 서로가 직접적인 영향을 주는지에 대한 인과관계는 정확하게 파악할 수 없다.

이는 회귀분석을 이용하면 가능하다.

- 상관 분석 : 변수 간의 관련성 분석
- 회귀 분석 : 변수 간의 인과관계 분석으로 사용범위가 넓음

예) 날씨가 더워질수록 아이스크림이 잘 팔린다.
날씨가 더워질수록 날파리가 늘어난다.
"아이스크림 판매가 늘어나니 날파리도 증가한다" 라고 판단할 수는 없다.

또 다른 변수인 "날씨"가 더워진 관계로 발생한
상관관계일 뿐이지 서로(아이스크림 / 날파리)가
직접적인 영향(인과관계)을 준 것은 아니다.
그러므로 임의의 변수 간에 관계를 파악하고
설명할 때는 신중을 기하는 것이 중요하다라고 생각된다.



상관관계 분석

상관분석은 기본적으로 변수가 2개 이상이므로 평균에서 치우침이 두 변수에 의해 발생하게 되기 때문에 분산 외에 공분산 값을 알아야 한다.
공분산은 두 개 이상의 확률변수에 대한 관계를 보여 주는 값이다.
즉, 확률변수 x와 y에 대해 x가 변할 때 y가 변하는 정도를 나타내는 값을 말한다.
관련이 없으면 0, 관련이 많을 수록 1에 가까워진다.

공분산을 표준화 시킨 것이 상관계수다. 두 확률변수의 공분산을 각각의 표준편차의 곱으로 나누어 준 것.
이렇게 되면 공분산의 영향력이 사라진다.

▶연구문제 예시

그렇다면, 상관분석은 어떤 연구문제에서 적용할 수 있을까?

연구문제 예시 - 변수 : 두 가지 연속 변수

1. 부모의 수입과 성적의 관련성
2. 키와 몸무게의 관련성
3. 나이와 스마트폰 사용시간의 관련성

위의 연구문제들은 **두 가지 연속 변수 간에 관련성이 있는지**를 알아보는데 관심이 있다.
두 가지 연속변수 간에 관련성이 있는지에 대한 문제를 검증하고자 한다면 상관분석을 적용할 수 있다.

상관계수 공식

$$r = \frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}}{\sqrt{\frac{\sum(x-\bar{x})^2}{n-1} \times \frac{\sum(y-\bar{y})^2}{n-1}}}$$

공분산

루트 빼면, 변수 x의 분산

루트 빼면, 변수 y의 분산

* Machine Learning (기계학습) *

• 사람과 기계와의 소통이 가능한 이유는 수 많은 알고리즘을 통해서 기계에게 학습을 시킨 후 새로운 데이터가 입력되면 기계 스스로가 해석할 수 있는 기계학습이 가능하기 때문이다.

• 알고리즘을 통해 컴퓨터에게 학습을 시킨 후에 새로운 자료가 입력된 경우 해당 자료의 결과를 분석하여 예측결과를 제공해 준다. 예를 들어 검색어 자동 완성, 악성코드 감지, 자료 인식 등의 예측을 필요로 하는 분야에서 사용될 수 있다.

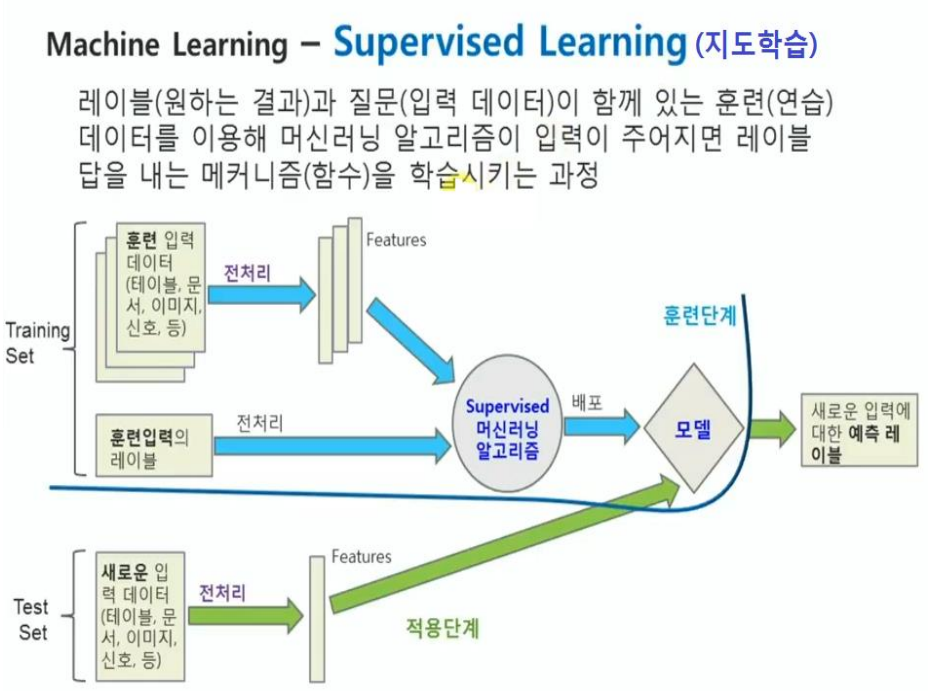
- Machine Learning?
  - 일상에서 접하는 Machine Learning
    - 상품의 추천/ 스팸 메일 분류/ 쿠폰발급/ 대출심사 등
  - 기업에서 적용하는 Machine Learning
    - Business decision / Productivity 증대/ Disease detection / Anomaly detection / Forecasting weather 등

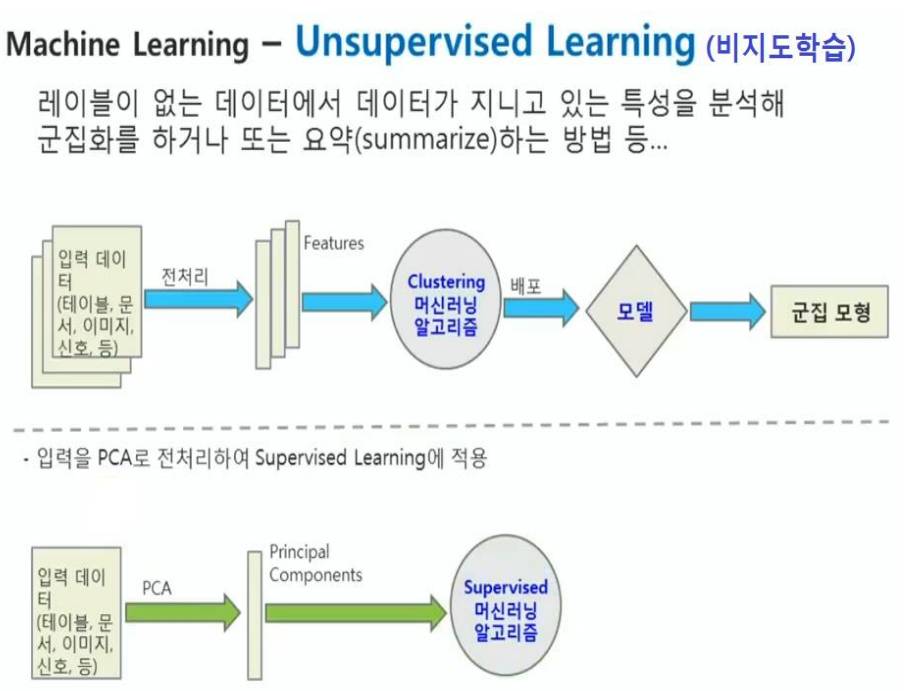
Machine Learning

*** 지도(Supervised)학습 / 비지도 (UnSupervised) 학습 비교 ***

분류	지도학습(교사학습)	비지도학습(비교사학습)
주관	사람이 개입	컴퓨터 자체
기법	확률과 통계 기반 추론 통계	패턴분석 기반, 데이터 마이닝
유형	회귀분석, 신경망 …	군집분석, 연관분석 …

* 지도학습은 독립변수와 종속변수가 있으나 비지도학습은 종속변수(label)가 없다.



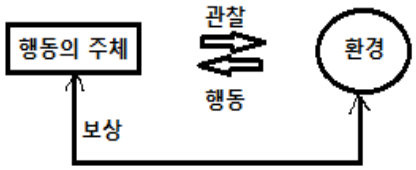


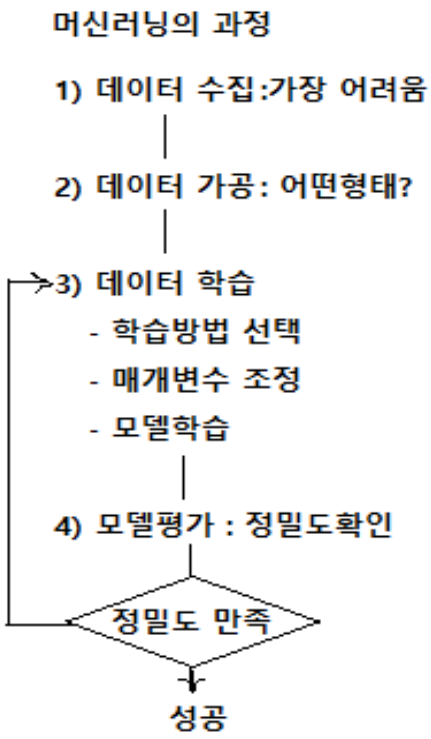
기계학습 방법	가능 분야	분석방법의 종류 및 알고리즘
지도학습 (Supervised Learning)	예측, 추정 (Prediction, Estimation)	✧ Linear Regression ✧ Regression Tree, Model Tree ✧ SVM(Support Vector Machine) ✧ Neural Network, Deep Learning ✧ ARIMA, Exponential Smoothing
	분류 (Classification)	✧ Decision Tree ✧ Logistic Regression, Discriminant Analysis ✧ k-NN(k-Nearest Neighbor), CBR(Case-Based Reasoning) ✧ Naïve Bayes Classification ✧ SVM, Neural Network ✧ Ensemble (Bagging, Boosting, Random Forest)
비지도학습 (Unsupervised Learning)	패턴/구조 발견 (Pattern/Rule)	✧ Association Rule Analysis, Sequence Analysis ✧ Network Analysis, Link Analysis, Graph theory ✧ Structural Equation Modeling, Path Analysis
	그룹화 (Grouping)	✧ k-Means Clustering, Hierarchical Clustering, Density-based Clustering, Fuzzy Clustering ✧ SOM(Self-Organizing Map)
	차원 축소 (Dimension Reduction)	✧ PCA(Principal Component Analysis), Factor Analysis, SVD(Singular Value Decomposition)
	영상,이미지,문자 (Video, Image, Text, Signal processing)	✧ Wavelet/Fast Fourier Transformation, DTW(Dynamic Time Warping), SAX(Symbolic Aggregate Approximation), Line/Circular Hough Transformation ✧ Text mining, Sentiment Analysis

강화학습(Reinforcement Learning)은 부분적으로 답을 입력하여 최적의 답을 찾는다. 이는 기계 학습의 한 영역으로 행동심리학에서 영감을 받았으며, 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법이다. 이러한 문제는 매우 포괄적이기 때문에 게임 이론, 제어 이론, 운용 과학, 정보 이론, 통계학, 시뮬레이션 최적화, 유전 알고리즘 등의 분야에서도 연구된다.

강화학습에서는 '행동의 주체'와 '환경(상황 또는 상태)'이 등장한다. 행동의 주체는 환경을 관찰하고, 이를 기반으로 의사결정을 내려 행동한다. 이 때 환경이 변화하면서 행동의 주체가 어떠한 보상을 받게 된다. 이를 기반으로 행동의 주체는 더 많은 보상을 얻을 수 있는 방향으로 행동을 학습하게 된다.

예) 교육학 – 스키너의 쥐 실험 : 강화이론 - 쥐는 최선의 행동을 통해 학습한 결과, 먹이를 얻을 수 있는 방법을 알 수 있게 됨.





머신러닝의 순서

1) 학습단계(Training) : 모델을 만들기 위해 데이터를 수집하고, 수집된 데이터에서 어떤 특징(feature)을 가지고 예측할 것인지 Feature들을 정의한 후에 이 feature를 기반으로 가설을 정의하고 학습시킨다.
2) 예측단계(prediction) : 학습이 끝나면 모델이 주어지고, 학습된 모델에 의해 결과값이 만들어진다.

머신러닝 응용분야

1) 클래스 분류 : 스펨메일 분류, 필기 인식, 증권사기 등
2) 클러스터링 : 사용자의 취향을 그룹화 : 광고 제공
3) 추천 : 구매상품에 대한 추천
4) 회귀 : 판매 예측, 주가변동 예측
5) 차원 축소

회귀분석(Regression Analysis)

변수 간의 인과관계를 밝히기란 매우 어려운 문제다. 수학적 방법 이외에 다양한 외적 조건도 따져봐야 한다. 회귀분석은 이런 과정 중에 하나에 불과하다.

- 특정변수(독립변수)가 다른 변수(종속변수)에 어떤 영향을 미치는가를 분석. 즉, 인과관계를 분석.
- 독립, 종속변수는 등간 또는 비율척도 (연속형 데이터)로 구성되어야 한다.
- 독립변수 중에서 종속변수에 영향을 미치는 변수를 규명하고, 이 들 변수들에 의해서 회귀방정식을 도출하여 회귀선을 추정한다. **Y = Wx + b**

  회귀분석에서 Weight와 bias의 최적값을 찾는 것(최소제곱법 사용)이 좋은 회귀식을 만들 수 있는 조건임.

- 회귀분석은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링 등의 통계적 예측에 이용될 수 있다.
- 회귀분석의 기본 가정 충족 조건으로 선형성, 잔차 정규성, 잔차 등분산성, 잔차 독립성, 다중 공선성 등.

회귀분석

~ **상관관계 분석 :** 변수 간의 **관련성을** 분석

~ **회귀분석 :** 변수 간의 **인과관계를** 분석하며, 사용범위가 넓은 분석방법이다. 이는 변수 들의 관련성 규명을 위해 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 데이터로부터 추정하는 통계방법이다.
독립변수의 값에 의해 종속변수의 값을 예측할 수 있다.

1) **단순회귀분석** : 독립변수와 종속변수가 각각 1개인 경우에 독립변수가 종속변수에 어떠한 영향을 미치는 가에 대한 인과관계를 분석.

*예) 연구모델 : 제품적절성(독립변수) -> 제품 만족도(종속변수)
제품의 품질과 가격수준을 결정하는 제품 적절성(독립변수)은 제품 만족도(종속변수)에 영향을 준다. 또는 영향을 주지 않는다. (영향이 있다 또는 없다.)

2) **다중회귀분석** : 여러 개의 독립변수로 1개의 종속변수에 미치는 영향 분석

*예) 연구모델 : 제품 적절성, 제품 친밀도 -> 제품 만족도
음료수 제품의 적절성과 친밀도는 제품 만족도에 영향을 미친다 또는 영향을 미치지 않는다.

회귀분석이란?

회귀분석은 **독립변인이 종속변인에 영향을 미치는지 알아보고자 할 때 실시하는 분석방법**이다.

연속형 자료에 따른 연속형 자료의 영향력을 검증하고자 할 때, **회귀분석**을 사용한다.

연속형 변수끼리 미치는 영향력이라고 하면 조금 헷갈릴 수 있다. 간단한 예를 통해 알아보겠다.

영향을 주는 변수	영향을 받는 변수	통계분석방법
연속형 자료	연속형 자료	회귀분석 구조방정식
	범주형 자료	로지스틱 회귀분석

연속형 변수로, 커피 맛, 가게 인테리어, 직원 친절도가 있다고 생각해보자.

위의 변수는 1점에서 5점 척도로 측정한다. 물론 7점 척도를 사용해도 된다.

커피 숍의 입장에서 위의 변수 중에 무엇이 만족도에 영향을 주는지 확인하고자 하는 것이다.

그리고 만족도를 높이려면, 무엇을 개선해야 하는지를 파악할 때, 회귀분석을 활용할 수 있다.

연구 문제 : A커피 숍의 커피의 맛, 가게 인테리어, 직원 친절도가 고객 만족도에 미치는 영향 통계분석 방법 : 회귀분석

연속형
A커피숍의 **{커피의 맛, 가게 인테리어, 직원 친절도}**가
{고객 만족도}에 미치는 영향
연속형

[독립변수] 연속형 자료 - 커피의 맛, 가게 인테리어, 직원 친절도

[종속변수] 연속형 자료 - 만족도

이처럼 두 연속형 자료가 미치는 영향에 대해 알아보고자 할 때 회귀분석을 사용한다.

▶회귀분석, 설문지 작성하기

예시를 통해, 알아보도록 하자. 회귀분석을 사용하는 연구문제에서는 설문지를 어떻게 작성해야 할까?

설문지 구성 예시

1. A사 커피브랜드의 다음 항목에 대한 만족도를 체크하십시오.

[커피의 맛] ① 매우 불만족 ② 불만족 ③ 보통 ④ 만족 ⑤ 매우 만족
[인테리어] ① 매우 불만족 ② 불만족 ③ 보통 ④ 만족 ⑤ 매우 만족
[직원 친절도] ① 매우 불만족 ② 불만족 ③ 보통 ④ 만족 ⑤ 매우 만족

2. A사 커피브랜드에 대해 전반적으로 얼마나 만족하십니까?

① 매우 불만족 ② 불만족 ③ 보통 ④ 만족 ⑤ 매우 만족

회귀분석은 영향을 주고 받는 변수들 모두 연속형 자료로 구성되어 있기 때문에,
이처럼, **연속형 자료를 얻을 수 있는 설문지를 구성**해야 한다.

▶회귀분석, 결과 분석하기

회귀분석의 경우에는 **통계수치 값을** 약간은 이해해야 한다.

R제곱과 F값이라는 것이 있는데, **R제곱은 식의 설명력이라고 보면 된다.**

독립변수 3개가 만족도를 얼마나 설명하느냐를 판단하고, 대략 30#정도 나오면 높은 수치라 할 수 있다.

F값은 모형 적합도를 나타낸다.

P값이 0.05보다 작으면 이 모형이 적합하다고 할 수 있다.
P값이 0.05보다 크면 이 모형은 부적합하다고 할 수 있다.

R제곱과 F값

R제곱 : 독립변수가 종속변수의 몇 퍼센트인가를 설명하는 수치

F값 : 회귀식의 적합도 ($p<0.05$보다 작아야 회귀식이 유의미함)

그 다음은 회귀식에 대해 분석하는 것이다.
B는 표준화되지 않은 영향력을 판단할 때 사용하고, β는 이 영향력의 상대적인 차이를 비교할 때 사용한다.

B와 β

B :비표준화 계수
절대적인 영향력의 크기

β :표준화 회귀계수
상대적인 영향력의 크기,
종속변수에 가장 큰 영향을 미치는 변수가 무엇인가를 판단할 때 활용

회귀식이 아래 예시처럼 나왔다고 생각해 보자.

회귀식 예시

만족도(y)=0.5 + 0.8*커피맛 + 0.7*인테리어 + 0.6*직원

➔ **커피맛이 1만큼 증가하면, 만족도는 0.8정도 증가한다고 예측할 수 있음**

회귀식을 통해, 이 값들이 얼마나 만족도에 영향을 미치는지 알 수 있으며,
커피 맛이 1이 증가하면, 만족도는 0.8정도가 증가한다고 볼 수 있다.

표준화, β 값 같은 경우는 커피 맛, 인테리어, 직원 친절도의 표준편차가 다르기 때문에,
자료가 얼마나 넓게 퍼지느냐 좁게 퍼지느냐에 따라 이것이 영향력이 큰지 작은지 판단할 수 있다.

여기에서는 커피 맛의 숫자가 가장 큰 것으로 볼 때, 커피 맛이 만족도에 제일 큰 영향을 줬다고 해석하면 된다.

▶연구문제 예시

회귀분석은 어떤 연구문제에서 적용할 수 있을까?

연구문제 예시

1. **부모의 수입** 이 **성적**에 미치는 영향
2. **키**가 **몸무게**에 미치는 영향
3. **나이**가 **스마트폰 사용시간**에 미치는 영향

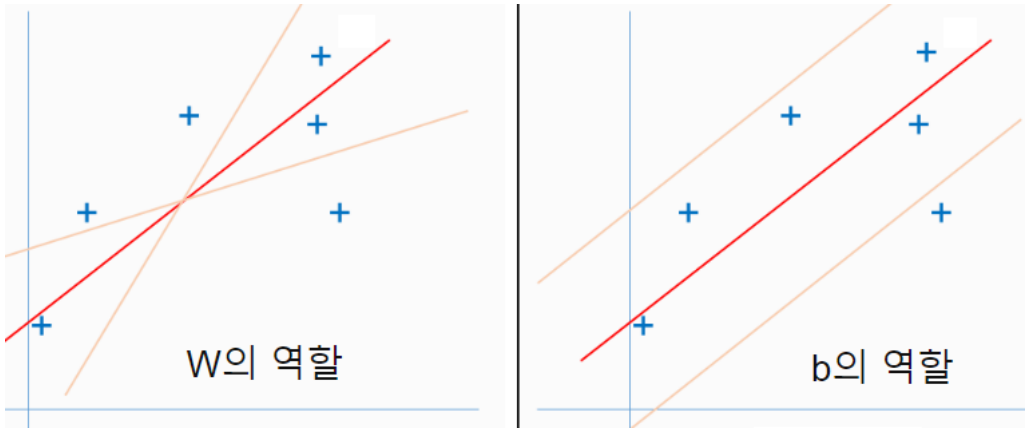
- 독립변수 : 연속변수
- 종속변수 : 연속변수

위의 연구 문제들은 **연속변수인 독립변인이 연속변수인 종속변인에 영향을 미치는지**를 알아 본다.

대부분의 연구는 독립변인이 종속변인에 영향을 미치는지를 검증하기 때문에,
학위논문에서 가장 많이 활용되는 것 중 하나가 바로 회귀분석 이다.

선형회귀분석에서 가설 함수 (Hypothesis function)

기존 데이터를 가장 잘 표현하는 직선을 결정하는 함수이며 이 직선은 W와 b를 변화시켜가면서 찾을 수 있다. 바로 이 최적의 W와 b를 찾는 것이 선형회귀분석의 목적이다.
그리고 이렇게 찾아진 W와 b를 함수에 대입하게 되면 기존의 데이터가 아닌 새로운 x값이 나타났을 때 그 새로운 x에 대한 y값을 예측할 수 있게 되는 것이다.
이 과정에서 가장 중요한 것은 기울기를 나타내는 W값이다. 즉, 전체적인 데이터가 그래프 상에서 어떤 모양(어떤 기울기의 직선형 그래프)과 가장 유사한가 하는 것을 확인하는 일이며 b는 그렇게 예측된 직선과 실제 데이터 간의 오차를 보정해준다고 생각하면 될 것이다.



선형회귀분석의 비용함수 (cost function)

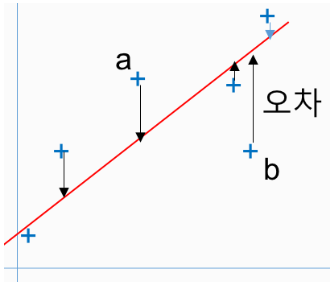
$$cost = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

선형회귀분석에서 가장 적절한 W와 b의 값을 찾는 과정에서 변화하는 W와 b의 값을 검증하는 함수.

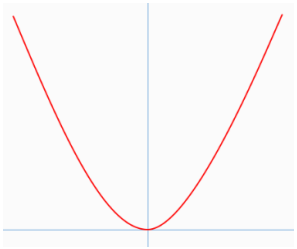
검증은 아주 상식적이며 단순하다. 우리가 예측할 직선상의 y값(가설함수의 결과인 y값)과 실제 좌표상의 y값(우리가 이미 알고있는 실제의y의 값)의 오차가 작으면 작을수록 좋은 값인 것이다.

그래서 우선 H(x(i)) - y(i)라는 공식이 도출된다(H(x(i))는 가설 함수의 결과를 의미한다).

즉, 이 비용 함수는 '오차'에 대한 함수인 것이다.



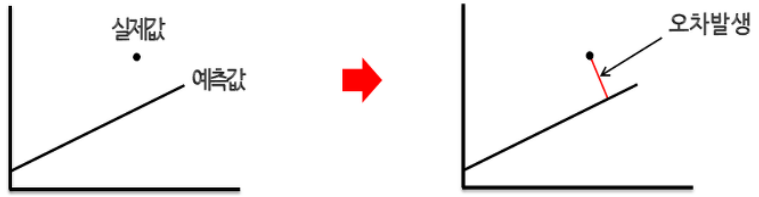
그런데 이 값을 그냥 사용한 것이 아니라 제곱을 했다.
이 것은 그냥 사용하게 될 경우 이 값은 음수와 양수로 그 부호가 달라진다.
그림에서 가설함수인 빨간 선 상의 H(x(i))값에서 a의 y값을
뺄 경우에는 그 결과가 음수가 나오고 b의 값을 뺄 경우에는 양수가 나온다.
각각의 데이터에 대해 이렇게 계산을 한 후 평균을 구하기 위해 합산을
하게 되면 양수와 음수가 상쇄되어 계산에 어려움이 생길 것이다.
따라서 제곱을 함으로써 양의 정수로 만들어주는 것이다.
여기까지 만들어진 공식이 (H(x(i)) - y(i))^2이다.
그리고 이렇게 제곱을 할 경우 결과 값이 크면 클수록 제곱 값은
기하급수적으로 커지기 때문에 일종이 패널티 역할을 하게 되는 것이다.

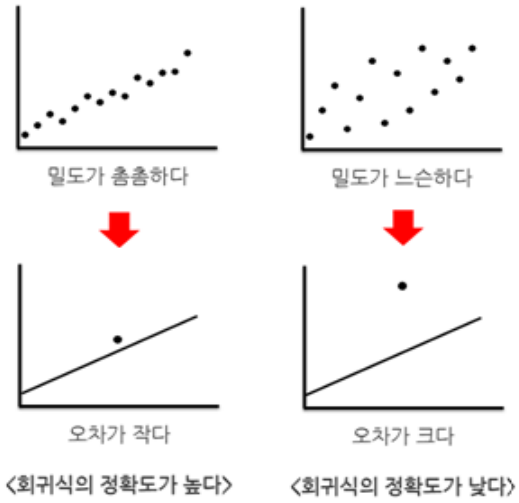


시그마(∑) 기호는 밑에 있는 i가 1부터 특정 수인 m까지 늘어나는 동안 기호 우측에 있는 계산식을 모두 더하라는 의미이고 수식의 가장 앞의 1/m은 m개의 계산 결과를 더한 것을 다시 m으로 나눈 것이니 바로 평균의 의미가 되는 것이다.
즉, cost 함수는 예측값과 실제 값의 차(오차)를 제곱한 모든 값의 평균이 되는 것이다.
이렇게 구한 cost 함수의 식을 그래프로 그려보면 곡선이 만들어진다. 즉, 이 함수는 제곱에 대한 함수이기 때문에 H(x(i)) - y(i)의 절대값이 크면 클수록 cost가 커지고 절대값이 작을 수록 cost가 0에 가까워지는 U자 형태의 그래프로 표현된다.

결정계수

결정계수는 상관분석의 상관계수와 유사하다. 회귀분석은 회귀식을 활용해서 무엇인가를 예측하는 분석이다. 그래서 무엇인가를 예측할 때, 회귀분석을 사용하면 눈대중으로 막 잡은 수치보다는 훨씬 신뢰할 수가 있다. 하지만 회귀분석으로 예측을 해도, 정답인 실제 값은 안 나온다. 다만 틀릴 확률이 존재하는 예측값이 나오면서, 항상 오차가 발생한다.



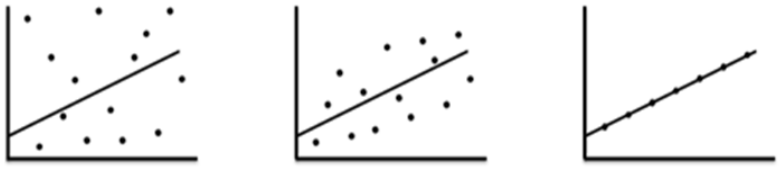


〈회귀식의 정확도가 높다〉 〈회귀식의 정확도가 낮다〉

점들이 모여 있는 밀도에 따라서, 오차의 크기가 다르다. 예를 들어 점들이 모여 있는 밀도가 촘촘할 경우에는, 예측값과 실제값이 얼마 차이 나지 않는다.(오차가 작다) 하지만 점들이 모여 있는 밀도가 느슨할 경우에는 예측값과 실제값이 많이 차이 난다.(오차가 크다) 그래서 똑같은 회귀분석이라도, 점들이 모여 있는 밀도에 따라 오차의 크기가 다르고, 그로 인해 회귀식의 정확도가 달라진다.

결정계수(R2)를 사용하면 회귀식이 얼마나 정확한지를 나타낼 수 있는데, 보통 숫자 0부터 1까지만(0 ≤ R2 ≤ 1) 사용한다.

- 결정계수가 0에 가까울수록 "회귀식의 정확도는 매우 낮다"고 할 수 있고,
- 결정계수가 1에 가까울수록 "회귀식의 정확도는 매우 높다"고 할 수 있다.
- 결정계수가 낮을수록 예측 값은 믿을 게 못되고, 높을수록 예측 값은 믿을 만하다.

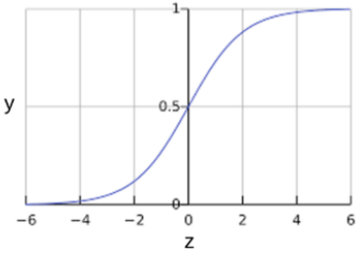


〈믿을 게 못 된다〉 〈어느 정도 믿을 만 하다〉 〈믿을 만 하다〉

로지스틱 회귀분석 (Logistic Regression)

종속변수와 독립변수 간의 관계로 예측모델을 생성한다는 점에서 선형회귀분석과 유사하다. 하지만 독립변수(x)에 의해 종속변수(y)의 범주로 분류한다는 측면에서 분류분석 방법이다. 분류 문제에서 선형 예측에 시그모이드 함수를 적용하여 가능한 각 불연속 라벨 값에 대한 확률을 생성하는 모델로 이진분류 문제에 흔히 사용되지만 다중클래스 분류(다중 클래스 로지스틱 회귀 또는 다항 회귀)에도 사용될 수 있다.

시그모이드 함수는 그림과 같다.



$$y = \frac{1}{1 + e^{-z}}$$

특징

1) 종속변수는 범주형이다. (이항형 : Yes/No 또는 다항형 : 예 - iris의 Spices 칼럼)
2) 정규성 : 정규분포 대신에 이항분포(성공확률이 P인 n회의 베르누이 독립시행) 를 따른다.
3) 로짓변환 : 종속변수의 출력범위를 0과 1로 조정하는 과정을 의미한다.
4) 활용분야 : 의료, 통신, 날씨 등 분류를 목적으로 다양한 분야에서 활용된다.

로지스틱 회귀분석은 종속변수(y 결과변수)가 범주형 데이터인 경우에 사용되는 기법으로 예측하고자 하는 것이 수치화 하기 힘든 변수, 예를 들어 어떤 고객이 부도를 낼 것인지의 여부, 타이타닉호에서의 생존 여부, 개인별 최종학력 알기 등의 경우에 사용할 수 있다.

참고 :

Odds는 성공 확률이 실행 확률에 비해 몇 배 더 높은 가를 나타낸다.

ex) 어떠한 사건이 일어날 확률을 p라고 하자. p (y = 1 | x)

로짓변환(Logit function) : Odds에 로그를 취한 함수로써 입력값의 범위가 [-∞ , +∞]일 때 출력 값의 범위를 [0, 1]]로 조정한다.

$$\text{odds ratio} = \frac{p(y=1|x)}{1-p(y=1|x)}$$

odds 비율 : 성공확률/실패확률

$$\text{logit}(p) = \log\frac{p}{1-p}$$

odd 비율을 로짓 변환

로지스틱 회귀분석

연속형 자료에 따른 범주형 자료의 영향력을 파악하기 위해, 로지스틱 회귀분석을 사용한다.

영향을 주는 변수	영향을 받는 변수	통계분석방법
연속형 자료	연속형 자료	회귀분석 구조방정식
	범주형 자료	로지스틱 회귀분석

로지스틱 회귀분석이라 하면, 통계를 좀 아는 분들도 어려워하지만, 사실 회귀분석과 큰 차이가 없다.

그럼 좀 더 구체적인 연구 문제 예시를 통해 알아보도록 하겠다.

연구문제 : 정치관심도, 여당선호도, 야당선호도가 선거참여에 미치는 영향
통계분석방법 : 로지스틱 회귀분석

연속형 자료 범주형 자료
{정치관심도, 여당선호도, 야당선호도}가 **{선거참여}**에 미치는 영향

[연속형 자료] 정치관심도, 여당선호도, 야당선호도

[범주형 자료] 선거참여 ➡ 선거 참여를 했다. ⇒ 1
선거 참여를 하지 않았다. ⇒ 0

▶로지스틱 회귀분석, 결과 분석하기

로지스틱 회귀분석 결과 값이 다음과 같이 나왔다고 생각해보자.

결과 예시	Exp(B)	P
정치 관심도	1.324	0.01
여당 선호도	0.800	0.01
야당 선호도	1.010	0.90

정치관심도에 대한 Exp(B) 값이 1.324 P 값이 0.001이다.

P 값이 0.05보다 작기 때문에 유의미하다는 결과가 나온다.
β 값은 0을 기준으로 0에 가까우면 유의미하지 않게 나오는데, **로지스틱 회귀분석은 1이 기준**이 된다.
해석할 때, 정치관심도가 1점 증가할수록, 선거에 참여할 확률이 1.324배 정도 높아진다는 결론이 나온다.
1배를 기준으로 해서 '몇 배 늘어난다'고 해석해야 한다.

만약에 **여당선호도에 대한 Exp(B)가 0.8, P 값이 0.01이 나온다**고 하면, 유의미한 것이다.

여당선호도가 1점 증가할수록 선거참여 할 확률은 0.8배로 떨어진다는 것이다.
* 0.8로 표시된 것은 0.8배 증가한 것이 아니라, 1배를 기준으로 20%정도 줄었다는 것이다.

결과 해석

정치 관심도가 1점 높아질수록, 선거참여 가능성은 1.324배 정도 높아진다.
여당 선호도가 1점 높아질수록, 선거참여 가능성은 0.800배 정도 낮아진다.
야당 선호도는 선거참여에 유의미한 영향을 미치지 못한다.

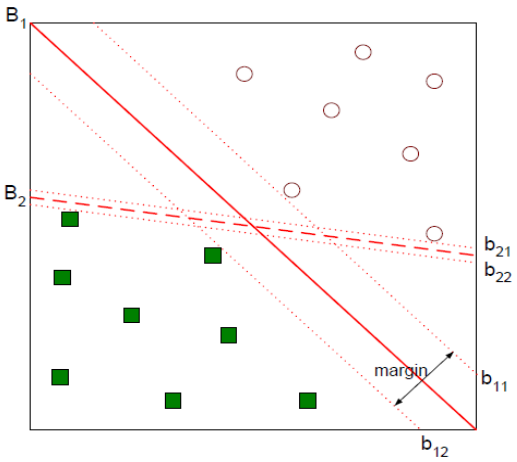
한마디로 여당을 좋아하는 사람들은 선거를 하지 않는다는 것이다.
로지스틱 회귀분석은 **연속형 변수들을 독립변수**로 놓고, **범주형 자료를 Yes or No로 나오는 범주형 변수를 종속변수로 활용**할 때, 진행할 수 있다.

Support Vector Machine(SVM)

분류와 회귀분석을 위해 주로 사용한다. 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형분류 모델을 만든다. 만들어진 분류 모델은 데이터가 사상된 공간에서 경계로 표현되는데 SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다. SVM은 선형 분류와 더불어 비선형 분류에서도 사용될 수 있다.

SVM은 다음과 같은 문제들을 해결하는데 주로 사용된다:

- SVM은 텍스트와 하이퍼텍스트를 분류하는데, 학습 데이터를 많이 줄일 수 있게 해 준다.
- 이미지를 분류하는 작업에서 SVM을 사용할 수 있다.
- SVM은 분류된 화합물에서 단백질을 구분하는 등의 의학 분야에 유용하게 사용된다.



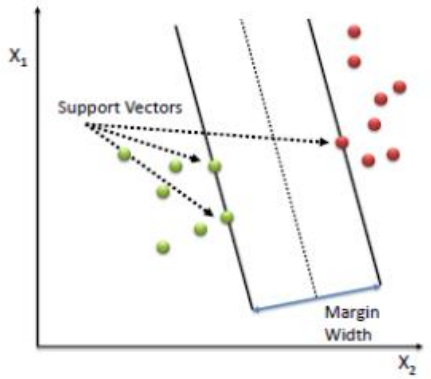
margin

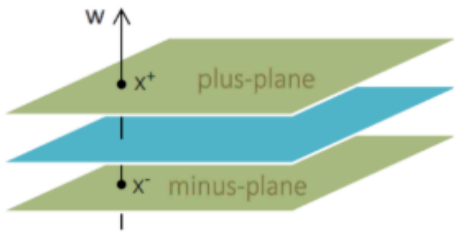
두 범주를 나누는 분류 문제를 푼다고 가정할 때 아래 그림에서 직선 B1과 B2 모두 두 클래스를 무난하게 분류하고 있음을 확인할 수 있다.

더 나은 분류경계면을 꼽으라면 B1일 것이다.

b12를 minus-plane, b11을 plus-plane, 이 둘 사이의 거리를 **마진(margin)**이라 한다.

SVM은 이 마진을 최대화하는 분류 경계면을 찾는 기법이다.





https://wikidocs.net/5719

Confusion Matrix (혼돈행렬, 혼동행렬)

분류모델의 예측 성공률, 즉 라벨과 모델의 분류 사이의 상관관계를 요약한 NxN 표다. 혼돈행렬의 축 중 하나는 모델이 예측한 라벨이고, 다른 축은 실제 라벨이다. N은 클래스의 수를 나타낸다. 이진분류 문제에서는 N=2이다. 예를 들어 다음은 이진 분류 문제에 대한 샘플 혼돈행렬이다.

	종양(예측)	비종양(예측)
종양(실제)	18	1
비종양(실제)	6	452

위 혼돈행렬에서는 실제로 종양이 있었던 샘플 19개 중 18개는 모델이 정확히 분류(참긍정 18개)했고, 1개는 종양이 없는 것으로 잘못 분류(거짓부정 1개)했다. 마찬가지로, 실제로 종양이 없었던 샘플 458개 중 452개는 정확히 분류(참음성 452개)되었고 6개는 잘못 분류(거짓양성 6개)되었다.

다중 클래스 분류 문제인 경우 혼돈행렬로 착오 패턴을 파악할 수 있다. 예를 들어 혼돈행렬은 필기 숫자를 인식하도록 학습된 모델이 4를 9로, 7을 1로 잘못 예측하는 경향이 있음을 드러낼 수 있다.

혼돈행렬은 알고리즘의 성능 평가지표로 많이 사용되며

정확성은 분류 모델 평가를 위한 측정항목 중 하나로 모델의 예측이 얼마나 정확한가를 보여준다.

정밀도 (precision, 양성으로 식별된 사례 중 실제로 양성이었던 사례의 비율은 어느 정도인가요?

모델이 **포지티브 클래스**를 정확히 예측한 빈도)

재현율 (recall, 실제 양성 중 정확히 양성이라고 식별된 사례의 비율은 어느 정도인가요?

가능한 모든 긍정 라벨 중에서 모델이 올바르게 식별한 것은 몇 개?)

등의 다양한 성능 측정항목을 계산하는 데 효과적인 정보를 포함한다.

정확도: TN + TP / 전체수 정밀도: TP / (TP + FP) 재현율: TP / (TP + FN)		Predicted (예측)	
		Positive	Negative
Observed (실제)	True	TP 참양성	FN 거짓음성
	False	FP 거짓양성	TN 참음성

http://bcho.tistory.com/1206

https://en.wikipedia.org/wiki/F1_score

분류: 참 대 거짓 (True, False) , 양성 대 음성 (Positive, Negative)

이솝 우화: 양치기 소년(일부)

어느 날 양치기 소년이 마을의 양떼를 돌보던 중 심심해졌습니다. 소년은 늑대가 보이지 않았지만 재미로 "늑대다!"라고 외쳤습니다. 마을 사람들은 양떼를 지키려고 뛰어왔지만 소년이 장난친 것을 알고는 몹시 화가 났습니다. ...
그러던 어느 날 밤, 양치기 소년은 양떼를 향해 다가오는 진짜 늑대를 보고 "늑대다!"라고 외칩니다. 하지만 마을 사람들은 소년이 또 거짓말을 한다고 생각해서 집에서 나오지 않습니다. 배고픈 늑대는 양들을 모두 잡아먹습니다.

이제 다음과 같이 정의해 보자.

"늑대다"는 양성 클래스, "늑대가 없다"는 네거티브 클래스.

'늑대 예측' 모델에서 발생할 수 있는 4가지 결과를 요약하면 2x2 혼돈행렬을 사용해 나타낼 수 있다.

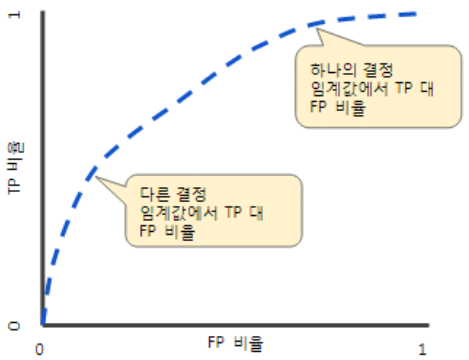
- 참양성은 모델에서 포지티브 클래스를 정확하게 평가하는 결과다.
- 참음성은 모델에서 네거티브 클래스를 정확하게 평가하는 결과다.
- 거짓양성은 모델에서 포지티브 클래스를 잘못 예측한 결과다.
- 거짓음성은 모델에서 네거티브 클래스를 잘못 예측한 결과다.

•참양성(TP):현실: 늑대의 위협이 있었다. •양치기의 외침: "늑대다." •결과: 양치기는 영웅! 예) 참 양성 결과수: 1	•거짓 음성(FN):현실: 늑대의 위협이 있었다. •양치기의 외침: "늑대가 없다." •결과: 늑대가 양들을 모두 잡아먹음. 예) 허위 음성 결과수: 8
•거짓 양성(FP):현실: 늑대의 위협이 없었다. •양치기의 외침: "늑대다." •결과: 마을 사람들은 양치기가 잠을 깨워서 화가 남. 예) 허위 양성 결과수: 1	•참 음성(TN):현실: 늑대의 위협이 없었다. •양치기의 외침: "늑대가 없다." •결과: 모두가 괜찮다. 예) 참 음성 결과수: 90

$$\text{정확성} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1 + 90}{1 + 90 + 1 + 8} = 0.91$$

$$\text{정밀도} = \frac{TP}{TP + FP} = \frac{1}{1 + 1} = 0.5 \qquad \text{재현율} = \frac{TP}{TP + FN} = \frac{1}{1 + 8} = 0.11$$

ROC 곡선 (Receiver Operating Characteristic curve, 수신자 조작 특성 곡선)

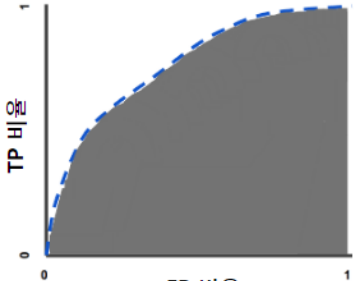


다양한 분류 임계값의 참 양성(TP) 및 허위 양성(FP) 비율

ROC 곡선(수신자 조작 특성 곡선)은
모든 분류 임계값에서 분류 모델의 성능을 보여주는 그래프로 두 매개변수를 표시한다.
- 참 양성 비율(TPR) : 재현율의 동의어이며 TPR=TPTP+FN 로 정의 된다.
- 허위 양성 비율(FPR) : FPR=FPFP+TN로 정의 된다.

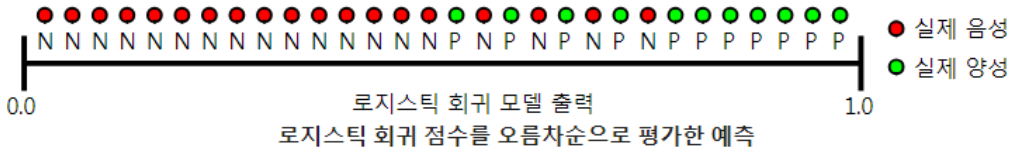
ROC 곡선은 다양한 분류 임계값의 TPR 및 FPR을 나타낸다.
분류 임계값을 낮추면 더 많은 항목이 양성으로 분류되어 거짓양성과 참양성이 모두 증가한다. 우측 그림에서는 일반 ROC 곡선을 보여준다.
ROC 곡선의 점을 계산하기 위해 분류 임계값이 다른 로지스틱 회귀 모형을 여러 번 평가할 수 있지만 이 방법은 효율적이지 않다.

다행히 이 정보를 제공할 수 있는 효율적인 정렬 기반 알고리즘이 있는데, 이를 AUC라고 한다.



AUC(ROC 곡선 아래 영역)

AUC(가능한 모든 분류 임계값을 고려하는 평가 측정항목으로 무작위로 선택한 긍정 예가 실제로 긍정일 가능성이 무작위로 선택한 부정 예가 긍정일 가능성보다 높다고 분류자가 신뢰할 확률이다.
'ROC 곡선 아래 영역'을 의미한다. 즉, AUC는 (0,0)에서 (1,1)까지 전체 ROC 곡선 아래에 있는 전체 2차원 영역을 측정한다(적분). AUC는 가능한 모든 분류 임계값에서 성능의 집계 측정값을 제공한다.
AUC를 해석하는 한 가지 방법은 모델이 임의 양성 예제를 임의 음성 예제보다 더 높게 평가할 확률이다.
예를 들어 다음 예에서는 로지스틱 회귀 예측의 오름차순으로 왼쪽에서 오른쪽으로 정렬되어 있다.



로지스틱 회귀 점수를 오름차순으로 평가한 예측

AUC는 임의의 양성(초록색) 예제가 임의의 음성(빨간색) 예제의 오른쪽에 배치되는 확률을 나타낸다.
AUC 값의 범위는 0~1 이다. 예측이 100% 잘못된 모델의 AUC는 0.0이고 예측이 100% 정확한 모델의 AUC는 1.0 이다.
AUC는 다음 두 가지 이유로 이상적이다.
- AUC는 **척도 불변이**다. AUC는 절대값이 아니라 예측이 얼마나 잘 평가되는지 측정한다.
- AUC는 **분류 임계값 불변이**다. AUC는 어떤 분류 임계값이 선택되었는지와 상관없이 모델의 예측 품질을 측정한다.

분류 알고리즘 중 나이브 베이즈 분류

머신러닝 알고리즘 중에서 분류 알고리즘으로 많이 사용되는 알고리즘으로 '나이브 베이즈'가 있다.

기계 학습분야에서, '나이브 베이즈 분류(Naïve Bayes Classification)'는 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종으로 광범위하게 연구되고 있다.

장점 : 소량의 데이터를 가지고 작업이 이루어지며, 여러 개의 분류 항목을 가질 수 있다.
예) 스팸메일 분류, 게시판 카테고리 분류 등

***조건부 확률 ***

P(B|A) : 사건 A에 대해서,사건 A가 발생했을 때 사건 B가 발생할 확률. "사건 A에 대한 사건 B의 조건부 확률"이라 한다. 예를 들어 어느 집단의 남학생일 확률이 P(A)라고 하고,학생의 키가 170이 넘는 확률을 P(B)라고 했을 때, 남학생 중에서 키가 170이 넘는 확률은 B의 조건부 확률이 되며 P(B|A)로 표현한다.

- **사전확률** :
  특정 사건이 일어나기 전의 확률로 베이즈 추론에서 관측자가 관측을 하기 전에 가지고 있는 확률분포를 말한다.

- **사후확률** :
  확률변수 관측에 대한 조건부 확률로, 어떤 사건이 발생하였고, 이 사건이 나온 이유가 무엇인지 P(B|A)란 식으로 나타낸 것이다. (B는관측한 사건. A는 B가 나올 수 있게 한 과거의 사실)

- **베이즈 정리** :
  두 확률변수의 사전확률과 사후확률 사이의 관계를 나타내는 정리다. 베이즈 정리는 사전확률로부터 사후확률을 구할 수 있는 개념이다.

참고 : http://nittaku.tistory.com/283

나이브 베이즈(Naive Bayes) - 예제 : 비오는 날

나이브 베이지안은 dataset의 모든 특징들이 동등하며 독립적이라고 가정한다. 예를 들어 사람들은 비가 오는 날에는 시간보다는 습도가 더 중요한 변수라고 생각할 수 있으나 나이브 베이지안 에서는 이런 사실을 무시하기 때문이다. 이런 가정에도 불구하고 분류학습에서 매우 정확한 결과 값을 유추할 수 있기 때문에 자주 사용되고 있다.
아래의 맑은 날과 비온 날 데이터를 이용해서 알아 나이브 베이즈에 대해 알아보자.

"만약 오늘 날씨가 좋고, 바람이 많이 불지 않고, 기압은 높은데, 온도가 낮다면 오늘은 비가 올 것인가? 안 올 것인가?"

위의 질문에 대해 식으로 바꿔보면

비가올 확률 =

$$\frac{P(\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도})}{P(\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도}) + P(\sim\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도})}$$

구해야 할 것은

$P(\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도})$, $P(\sim\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도})$

비와 다른 변수들 간의 관계 (돗수분포표)

	날씨가 좋은가?		바람이 많이 부는가?		기압이 높은가?		온도가 높은가?		
	Yes	No	Yes	No	Yes	No	Yes	No	계
비 온날	2	6	6	2	8	0	5	3	8
안온날	8	4	2	10	2	10	6	6	12
계	10	10	8	12	10	10	11	9	20

비와 변수들 간의 관계

	날씨가 좋은가?		바람이 많이 부는가?		기압이 높은가?		온도가 높은가?		
	Yes	No	Yes	No	Yes	No	Yes	No	계
비 온날	2/8	6/8	6/8	2/8	8/8	0/8	5/8	3/8	8
안온날	8/12	4/12	2/12	10/12	2/12	10/12	6/12	6/12	12
계	10	10	8	12	10	10	11	9	20

결론 : 날씨가 좋으며, 바람이 많이 불지 않고, 기압은 높은데 온도가 낮다면 비가 올 확률은 2.7%이고 비가 안 올 확률은 97.3% 로 계산된다. 이 결과에 따르면 날씨가 좋고, 바람이 안불고 기압이 높은데,온도가 낮은 날은 비가 안 올 가능성이 높은 것으로 판단된다.

$$P(\text{비} \mid \text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})$$

$$= \frac{P(\text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도} \mid \text{비})P(\text{비})}{P(\text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})}$$

$$= \frac{P(\text{날씨}|\text{비})\ P(\sim\text{바람}|\text{비})\ P(\text{기압}|\text{비})\ P(\sim\text{온도}|\text{비})P(\text{비})}{P(\text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})}$$

$$\fallingdotseq P(\text{날씨}|\text{비})\ P(\sim\text{바람}|\text{비})\ P(\text{기압}|\text{비})\ P(\sim\text{온도}|\text{비})P(\text{비})$$

1. $P(\text{날씨}|\text{비})\ P(\sim\text{바람}|\text{비})\ P(\text{기압}|\text{비})\ P(\sim\text{온도}|\text{비})P(\text{비})$
   $= (2/8) * (2/8) * (8/8) * (3/8) * (8/20) = 0.009375$
2. $P(\text{날씨}|\sim\text{비})\ P(\sim\text{바람}|\sim\text{비})\ P(\text{기압}|\sim\text{비})\ P(\sim\text{온도}|\sim\text{비})P(\sim\text{비})$
   $= (8/12) * (10/12) * (2/12) * (6/12) * (12/20) = 0.33333$

비가 올 확률

$$= \frac{P(\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도})}{P(\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도}) + P(\sim\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도})}$$

$$= \frac{0.009375}{0.009375+0.333333} = 0.027 \rightarrow 2.7\%$$

비가 오지 않을 확률

$$= \frac{P(\sim\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도})}{P(\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도}) + P(\sim\text{비} \mid \text{날씨}\cap\sim\text{바람}\cap\text{기압}\cap\sim\text{온도})}$$

$$= \frac{0.333333}{0.009375+0.333333} = 0.973 \rightarrow 97.3\%$$

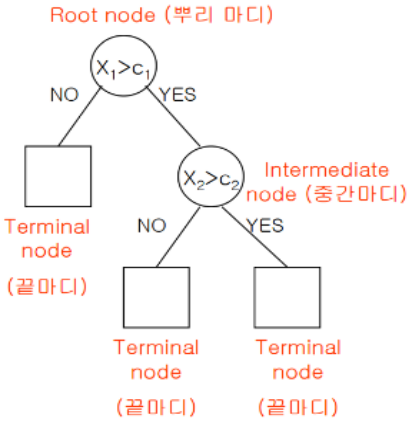
분류분석 - Decision Tree -

- Decision Tree는 여러 가지 규칙을 순차적으로 적용하면서 독립 변수 공간을 분할하는 분류 모형이다.

  분류(classification)와 회귀 분석(regression)에 모두 사용될 수 있다. 해석이 쉽다.

- 비모수 검정 : 선형성, 정규성, 등분산성 가정 필요 없음

- 단점 : 유의수준 판단 기준 없음(추론 기능 없음), 비연속성/ 선형성 또는 주효과 결여/ 비안정성(분석용 자료에만 의존하므로)으로 새로운 자료의 예측에서는 불안정할 수 있음.



의사결정나무를 일반화한 그림은 아래와 같다.

전체적으로 보면 나무를 뒤집어 놓은 것과 같은 모양이다. 초기 지점은 root node 이고 분기가 거듭될 수록 그에 해당하는 데이터의 개수는 줄어든다. 각 terminal node에 속하는 데이터의 개수를 합하면 root node의 데이터 수와 일치한다. 바꿔 말하면 terminal node 간 교집합이 없다는 뜻이다. 한편 terminal node의 개수가 분리된 집합의 개수이다. 그림처럼 terminal node가 3개라면 전체 데이터가 3개의 부분집합으로 나눠진 셈이다.

*** 의사결정나무의 형성과정**

1) 알고리즘: CHAID, CART, C45. 분석의 목적과 자료구조에 따라 적절한 분리기준(split criterion)과 정지규칙(stopping rule)을 지정하여 의사결정나무를 얻는다.
2) 가지치기: 분류오류(classification)를 크게 할 위험(risk)이 높거나 부적절한 추론규칙(induction rule)을 가지고 있는 가지(branch)를 제거.
3) 타당성 평가: 이득도표(profit chart)나 위험도표(risk chart)와 같은 모형평가 도구 또는 평가용 데이터(validation data)에 의한 교차타당성(cross validation) 등을 이용하여 의사결정나무를 평가
4) 해석 및 예측: 의사결정나무를 해석하고 예측 모형을 구축한다.

분류분석 - Decision Tree -

*** 분류나무(classification tree): 이산형 목표변수의 경우**

목표변수의 각 범주에 속하는 빈도(frequency)에 기초하여 분리가 일어남.

사용되는 분리기준
- 카이제곱 통계량의 p-값 (p-value of Chi Square statistic)
- 지니 지수(Gini index)
- 엔트로피 지수(Entropy index)

*** 회귀나무(regression tree): 연속형 목표변수의 경우**

목표변수가 연속형(구간형)인 경우 목표변수의 평균(mean)과 표준편차(standard deviation)에 기초하여 마디의 분리가 일어난다.

*** 정지규칙과 가지치기**

정지규칙(stopping rule) – 더 이상 분리가 일어나지 않고 현재의 마디가 끝 마디가 되도록 하는 여러 규칙.

지나치게 많은 마디는 –> 새로운 자료 적용 -> 예측 오차 높아짐. -> 가지치기(pruning) 필요.

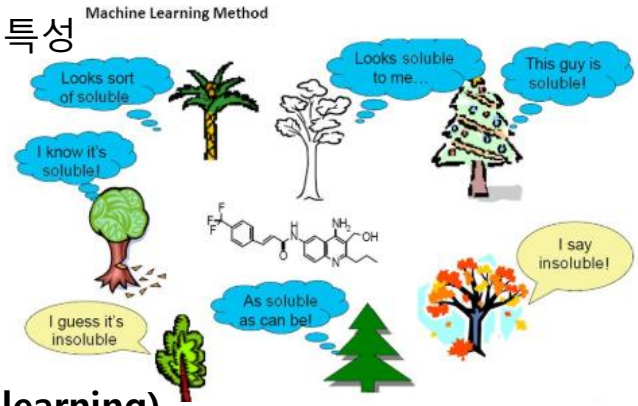
*** 분석 사례 – 신용평가 문제**

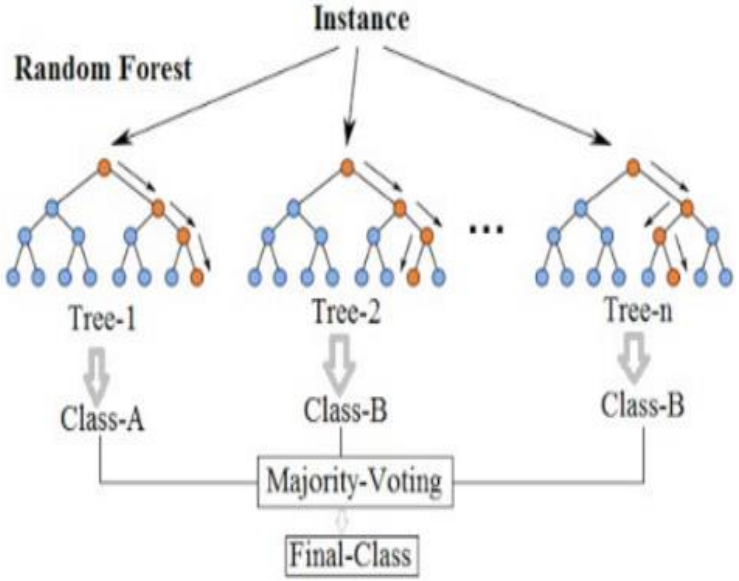
한 은행의 신용평가 부서에서는 대출 승인에 대한 의사결정 과정을 자동화하기 위해서 각 고객에 대한 신용평가 지수(credit score) 모형을 만들고자 한다.
이를 위해 데이터 셋을 구성하였으며, 생성된 모형은 대출승인 여부를 결정하는 예측모형으로 사용될 것이다. 그러나 대출이 거절된 고객에게는 그 사유를 설명할 수 있어야 하므로 연구자가 모형을 충분히 이해할 수 있어야 한다. 따라서 적절한 예측력과 충분한 설명력을 확보하기 위해 의사결정나무를 이용하여 모형화를 시도해야 한다.

랜덤 포레스트 (Random Forest) 분류모형

랜덤 포레스트는 분류, 회귀분석 등에 사용되는 앙상블 학습방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로 부터 분류 또는 평균 예측치(회귀 분석)를 출력함으로써 동작한다.

- Decision Tree에 비해 높은 정확성
- 간편하고 빠른 학습 및 테스트 알고리즘
- 변수 소거 없이 수천 개의 입력 변수들을 다루는 것이 가능
- 임의화를 통한 좋은 일반화 성능
- 다중 클래스 알고리즘 특성





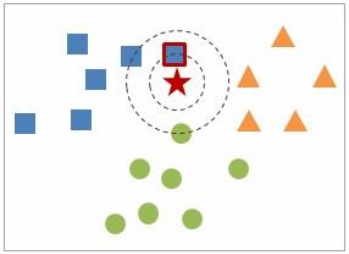
*** 앙상블 기법(ensemble learning)**

구축한 트리에는 랜덤성이 없는데 어떻게하면 랜덤하게 트리를 얻을 수 있나? 라는 의문이 든다.

랜덤포레스트에서는 데이터를 bootstrap 해서 포레스트를 구성한다. 이를 bootstrap aggregating 또는 begging 이라고 하는데, 전체 데이터를 전부 이용해서 학습시키는 것이 아니라 샘플의 결과물을 각 트리의 입력 값으로 넣어 학습하는 방식이다. 이러면 각 트리가 서로 다른 데이터로 구축되기 때문에 랜덤성이 생기게 된다. 그리고 파티션을 나눌 때 변수에 랜덤성을 부여한다. 즉, 남아있는 모든 변수 중에서 최적의 변수를 선택하는 것이 아니라 변수 중 일부만 선택하고 그 일부 중에서 최적의 변수를 선택하는 것이다.

K-NN (K-Nearest Neighber)

레이블이 있는 데이터를 사용하여 분류 작업을 하는 알고리즘이다. 데이터로부터 거리가 가까운 k개의 다른 데이터의 레이블을 참조하여 분류한다. 대개의 경우에 유클리디안 거리 계산법을 사용하여 거리를 측정하는데, 벡터의 크기가 커지면 계산이 복잡해진다.



그림을 참고하자. K-NN은 새로 들어온 " ★은 ■ 그룹의 데이터와
가장 가까우니 ★은 ■ 그룹이다. " 라고 분류하는 알고리즘이다.
여기서 k의 역할은 몇 번째로 가까운 데이터까지 살펴볼 것인가를 정한 숫자다.

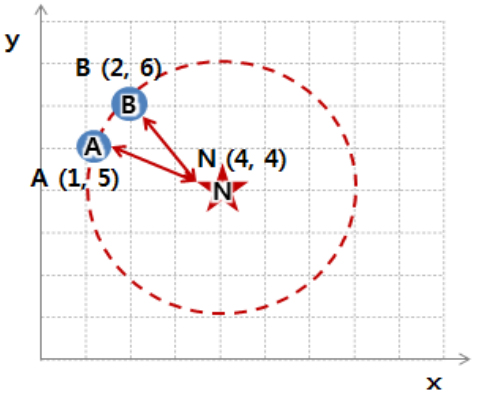
- 거리척도의 단위문제 – 표준화 필요 – 달러 대 원화 (유클리드 거리)

k를 정하기 전에 선행되어야 하는 작업이 있다. 바로 표준화. K-NN에서 가깝다는 개념은 유클리드 거리(Euclidean Distance)로 정의하는데, 유클리드 거리를 계산할 때는 단위가 매우 중요하다.

유클리드의 거리는 아래와 같이 계산한다.

$$\sqrt{(Ax - Bx)^2 + (Ay - By)^2}$$

오른쪽 그림에서 A-N 간의 유클리드 거리는 3.162 이고
B-N 간의 유클리드 거리는 2.828 로, B가 더 가깝다.



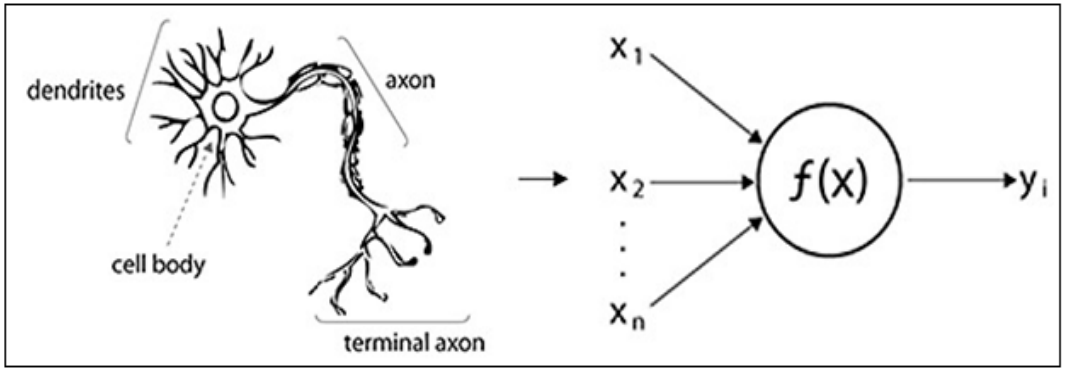
- **장점** : 알고리즘이 간단해 구현이 쉽다. 수치 기반 데이터 분류 작업에서 성능이 좋다 .
- **단점 :** 학습 데이터의 양이 많으면 분류 속도가 느려진다. 차원(벡터)의 크기가 크면 계산량이 많아진다

Neural Network

사람이 뇌를 사용해 문제를 처리하는 방법과 비슷한 방법으로 문제를 해결하기 위한 알고리즘을 사용한다. 사람은 뇌의 기본 구조 조직인 뉴런(neuron)과 뉴런이 연결되어 일을 처리하는 것처럼, 수학적 모델로서의 뉴런이 상호 연결되어 네트워크를 형성할 때 이를 신경망이라 한다.

생물학적 신경망을 모방하여 인공신경망을 모델링한 내용을 보면 처리 단위 측면에서는 생물적인 뉴런(neurons)이 노드(nodes)로, 연결성(Connections)은 시냅스(Synapse)가 가중치로 모델링 되었다.

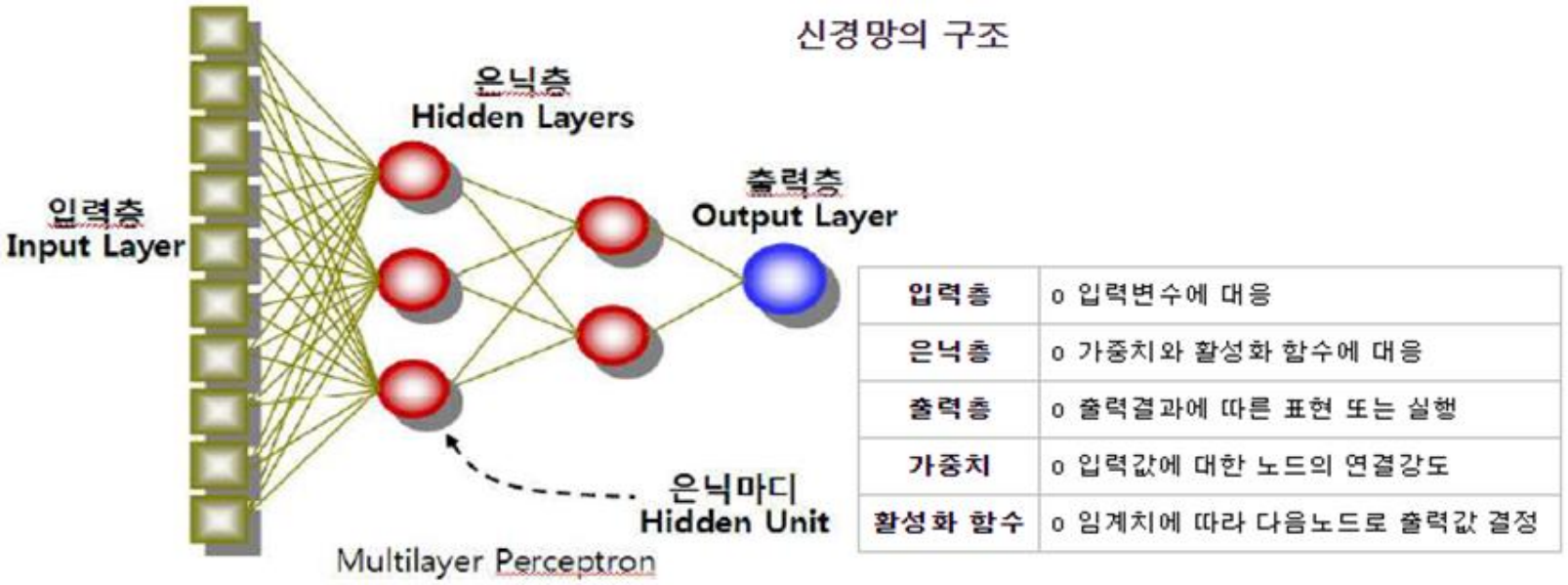
생물학적 신경망	인공신경망
세포체	노드(Node)
수상돌기(dendrite)	입력(Input)
축삭(Axon)	출력(Output)
시냅스	가중치(Weight)



생물학적인 신경망과 구별하여 특히 인공 신경망(artificial neural network)이라고도 한다. 신경망은 각 뉴런이 독립적으로 동작하는 처리기의 역할을 하기 때문에 병렬성(parallellism)이 뛰어나고, 많은 연결선에 정보가 분산되어 있기 때문에 몇몇 뉴런에 문제가 발생하더라도 전체 시스템에 큰 영향을 주지 않으므로 결함 허용(fault tolerance) 능력이 있으며, 주어진 환경에 대한 학습 능력이 있다.

신경망 분석의 장단점

장점	적용 가능한 문제의 영역이 넓음	o 입력,출력마디에 이산형,연속형 변수 모두 사용가능하며 기법을 적용할 수 있는 문제의 영역이 Decision Tree나 통계에 비해 넓음
	제품이 많음	o 상용화된 데이터마이닝 제품이 많으며, 제품 선택의 폭이 넓음
단점	과정에 대한 설명부족	o 분류나 예측결과만 제공할 뿐 결과에 대한 근거를 설명하지 못함
	모형 구축의 어려움	o 복잡한 학습과정을 거치기 때문에 모형 구축시 많은 시간이 소요. 따라서 입력변수의 수가 너무 많으면 통계나 Decision Tree를 이용, 변수 선별 후 구축하는 대안을 고려 할 수 있음
	전문가 필요	o 다양한 Parameter값을 설정하는 작업이 전문성을 필요로 하기 때문에 비전문가들이 사용하기 어려움

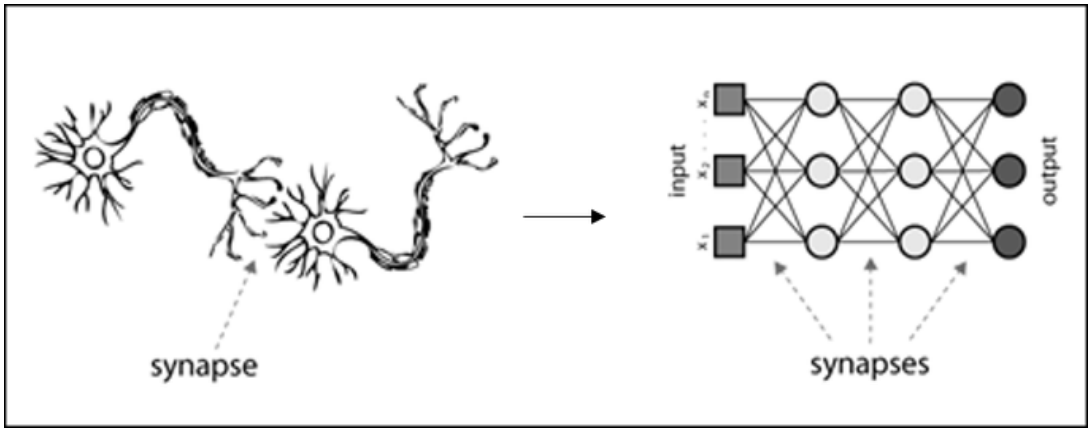


입력층	o 입력변수에 대응
은닉층	o 가중치와 활성화 함수에 대응
출력층	o 출력결과에 따른 표현 또는 실행
가중치	o 입력값에 대한 노드의 연결강도
활성화 함수	o 임계치에 따라 다음노드로 출력값 결정

Neural Network

신경세포(뉴런)의 입력은 다수이고 출력은 하나이며, 여러 신경세포로부터 전달되어 온 신호들은 합산되어 출력된다. 합산된 값이 설정값 이상이면 출력 신호가 생기고 이하이면 출력 신호가 없다.

- 연결(connection) : Synapse vs. weight



인간의 생물학적 신경세포가 하나가 아닌 다수가 연결되어 의미 있는 작업을 하듯, 인공신경망의 경우도 개별 뉴런들을 서로 시냅스를 통해 서로 연결시켜서 복수개의 계층(layer)이 서로 연결되어 각 층간의 연결강도는 가중치로 수정(update) 가능하다. 이와 같이 다층 구조와 연결강도로 학습과 인지를 위한 분야에 활용된다.

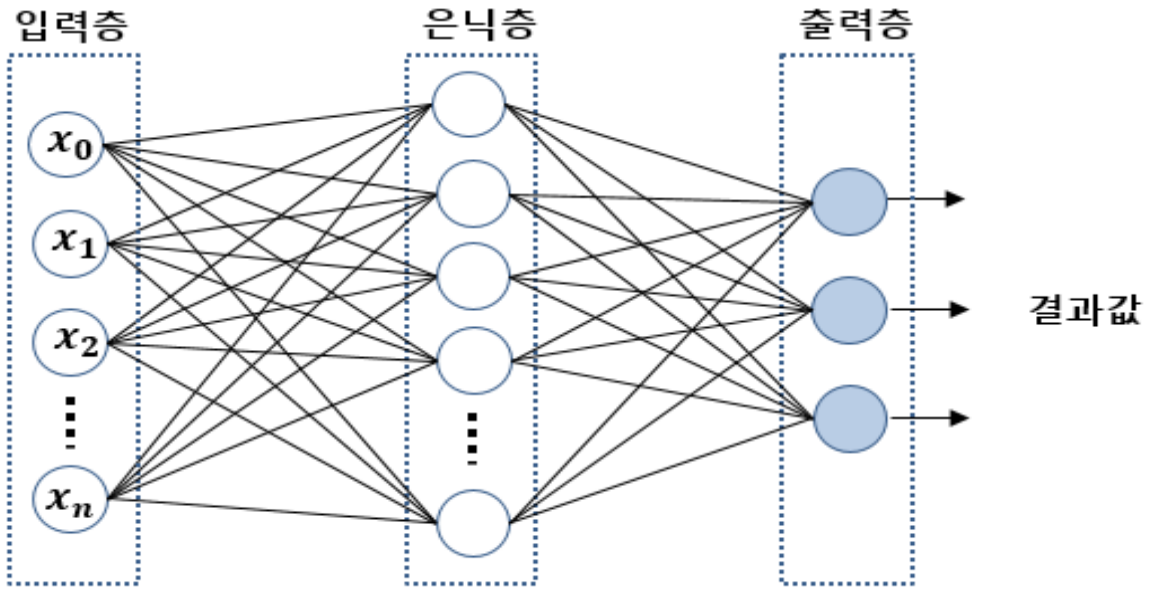
인공 신경망은 다음과 같은 작업에서 사용될 수 있다.

: 함수 추론, 회귀 분석, 시계열 예측, 근사 모델링, 패턴 인식 및 순서 인식 그리고 순차 결정 같은 분류 알고리즘, 필터링, 클러스터링, 압축 등의 데이터 처리, 인공 기관의 움직임 조정 같은 로봇 제어, 컴퓨터 수치 제어 등

다층 신경망 (MLP)

인공신경망인 단층 퍼셉트론은 그 한계가 있는데, 비선형적으로 분리되는 데이터에 대해서는 제대로 된 학습이 불가능하다는 것입니다. 예를 들면 단층 퍼셉트론으로 AND연산에 대해서는 학습이 가능하지만, XOR에 대해서는 학습이 불가능하다는 것이 증명되었습니다.

이를 극복하기 위한 방안으로 입력층과 출력층 사이에 하나 이상의 중간층을 두어 비선형적으로 분리되는 데이터에 대해서도 학습이 가능하도록 다층 퍼셉트론(줄여서 MLP)이 고안되었습니다. 아래 그림은 다층 퍼셉트론의 구조의 한 예를 보인 것입니다.



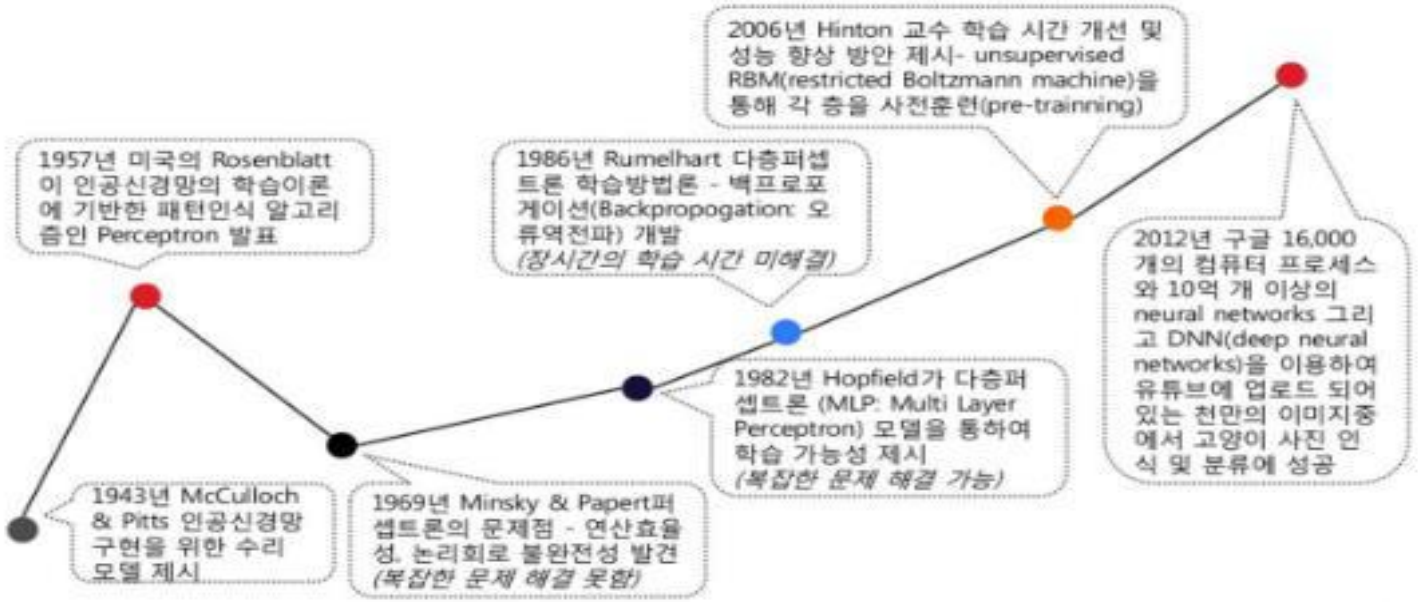
입력층과 출력층 사이에 존재하는 중간층을 숨어 있는 층이라 해서 은닉층이라 부릅니다. 입력층과 출력층 사이에 여러개의 은닉층이 있는 인공 신경망을 **심층 신경망(deep neural network)**이라 부르며, 심층 신경망을 학습하기 위해 고안된 특별한 알고리즘들을 **딥러닝(deep learning)**이라 부릅니다.

Deep Learning

다층 신경망의 성능문제의 해결방법이 2006년 Hinton의 "A fast learning algorithm for deep belief nets"를 통해 제시되면서, 딥러닝으로 주목 받음.

딥러닝은 은닉층을 계산에 활용하는 방식에 따라 Convolutional Neural Network, Deep Belief Network, Recurrent Neural Network 등으로 나뉨.

충분한 성능을 얻기 위해서는 대용량 계산의 처리가 가능해야 하며, GPU 또는 클라우드 컴퓨팅 환경의 사용이 필요하다.



비지도 학습 - 군집분석(Clustering)

Clustering?

– Cluster의 개수나 구조에 관한 특별한 사전 가정 없이 개체들 사이의 유사성/거리에 근거해 cluster를 찾고 다음 단계의 분석을 하게 하는 기법.

– 유사한 개체들을 cluster로 그룹화하여 각 집단의 성격을 파악.

Clustering 장점

– 데이터에 대해 탐색적 기법으로, 데이터 내부구조 등이 주어지지 않아도 자료구조를 탐색

– 추가적인 분석을 위해 사용할 수 있음.

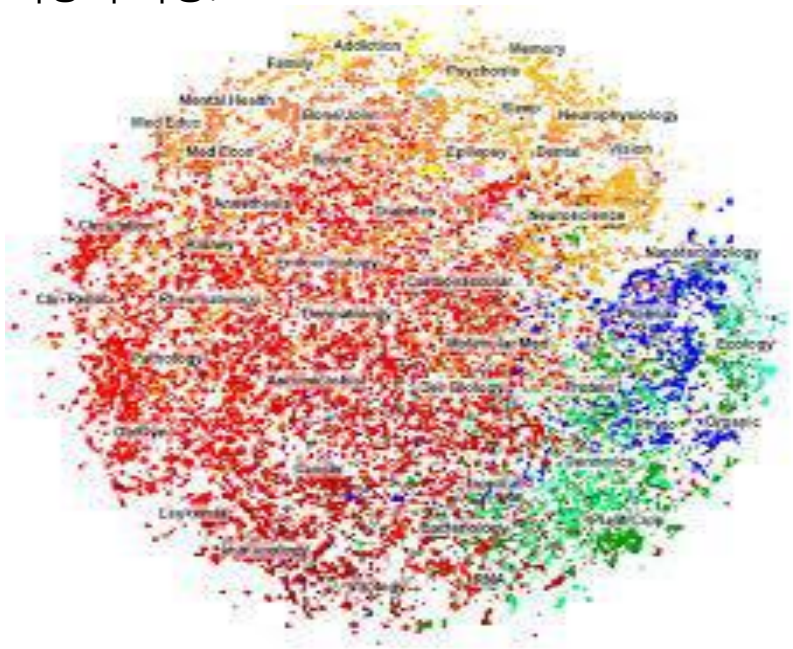
– 유사성, 비유사성 만 계산할 수 있다면 여러 형태 데이터 적용이 가능.

Clustering 단점

– 자료 유형이 혼합된 경우, 거리정의 등이 어려울 수 있음

– 초기 군집 수 설정이 중요

– 결과 해석에 주의



군집분석이란?

데이터 간의 유사도를 정의하고 그 유사도에 가까운 것부터 순서대로 합쳐 가는 방법으로,
유사도의 정의에는 거리나 상관계수 등 여러가지가 있다.

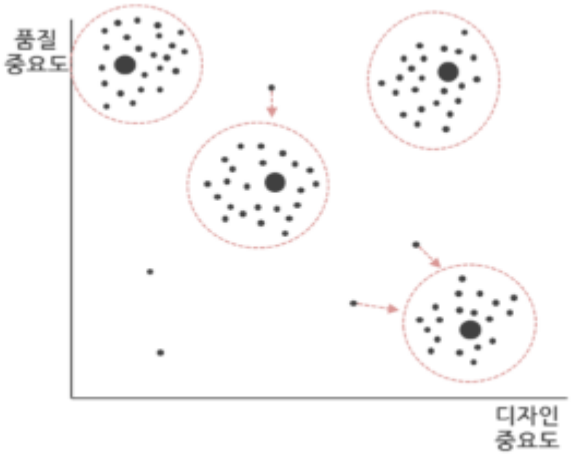
군집분석은 익숙하지 않아서 많이 어려워하지만, 개념만 제대로 이해하면 그리 어려운 통계분석 방법은 아니다.

군집분석은 본 분석을 들어가기 앞서, 사전작업이라고 볼 수 있다.
비슷한 특성을 가진 개체를 그룹으로 만들고, 그 그룹 간 서로 비교하는 것이 일반적인 통계분석의 흐름이다.

그럼 조금 더 구체적인 예시와 함께 알아보도록 하겠다.
예를 들어, 소비자의 특성을 군집(그룹)을 분리한다고 생각해보자.

다양한 변수가 있다면 3,4차원의 그래프로 표현될 수 있지만,
X축의 디자인 중요도, Y축의 품질 중요도로 2차원 그래프로 예를 들어보겠다.

어떤 제품을 볼 때, 품질을 보는 사람이 있는지, 디자인을 중요하게 보는 사람이 있는지 본다고 가정하겠다.
디자인만 중요시 하는 사람도 있고, 품질을 중요시 하는 사람도 있을 거고,
품질과 디자인 둘 다를 중요시 하는 사람도 있을 것이다.



그룹을 가장 가까이 묶인 것들끼리 그룹이 나뉠 수 있다.
이렇게 군집을 나누면, 특성을 4개의 그룹으로 분리할 수 있다.

군집분석

- 개인 또는 여러 개체를 유사한 속성을 지닌 대상들 끼리 그룹핑 하는 탐색적 다변량 분석기법이다.

- 거리값(Distance Measure)을 이용해 가까운 거리에 있는 것들끼리 묶어 분류한다.

- 계층적 군집분석과 비계층적 군집으로 분류할 수 있다.

1) 계층적 군집분석 :

개별 대상 간의 거리에 의하여 가장 가까이 있는 대상들로 부터 시작하여 결합해 감으로써 나무모양의 계층적 구조를 형성해 나가는 방법으로 이 과정에서 군집의 수가 감소한다. 계층적 군집분석은 군집이 형성되는 과정을 정확하게 파악할 수 있다는 장점이 있으나 자료의 크기가 크면 분석하기 어렵다는 단점이 있다.

방법 : 단일결합법, 완전결합법, 평균결합법, 중심결합기준법, Ward법

2) 비계층적 군집분석 :

군집의 수를 정한 상태에서 설정된 군집의 중심에서 가장 가까운 개체를 하나씩 포함해 나가는 방법으로 많은 자료를 빠르고 쉽게 분류할 수 있지만 군집의 수를 미리 정해 줘야 하고 군집을 형성하기 위한 초기값에 따라 군집의 결과가 달라진다는 어려움이 있기 때문에 계층적 군집분석을 통해 대략적인 군집의 수를 파악하고 이를 초기 군집 수로 설정한다.

방법: k-means clustering

K-means Clustering

군집화는 아무런 정보가 없는 상태에서 데이터를 분류하는 방법이다. K-means Clustering 이란 데이터 분류 종류를 K개 라고 했을 때 입력한 데이터 중 임의로 선택된 K 개의 기준과 각 점들의 거리를 오차로 생각하고 각각의 점들은 거리가 가장 가까운 기준에 해당한다고 생각하는 것이다. 그리고 이제 각각 기준에 해당하는 점들 모두의 평균을 새로운 기준으로 갱신해 나가게 된다. 이렇게 해서 가장 적절한 중심점들을 찾는 것이다. 이렇게 학습을 반복하면 데이터를 분류할 수 있게 된다.

클러스터링은 보통 4개의 유형으로 구분된다

- 클러스터 중심(centroid) 또는 평균 기반 클러스터링 k-means
- 빈도수가 많은 중간점(medoid) 기반 클러스터링 k-medoids
- 계층적 클러스터링
- 밀도 기반 클러스터링

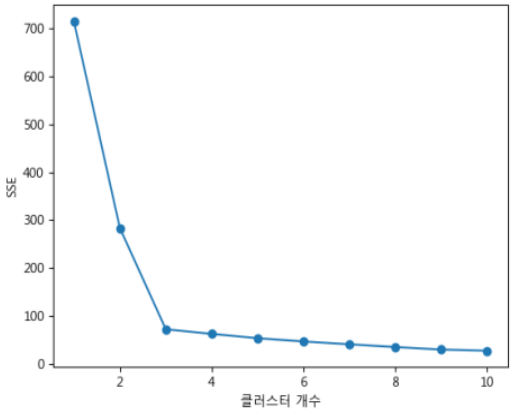
KNN은 가까운 데이터는 같은 label일 가능성이 크다고 가정하고 새로운 데이터가 들어오면, 그 데이터와 가장 가까운 k개의 데이터를 training set에서 뽑는다. 뽑은 k개의 데이터들의 label을 관측하고 그 중 가장 많은 label을 새로운 데이터의 label로 assign하는 알고리즘이다 (이런 방식을 majority voting이라고 한다). 이때 '가까움'은 Euclidean distance로 측정해도 되고, 다른 metric이나 measure를 사용해도 된다. 이때 distance 혹은 similarity를 측정하기 위해서 반드시 metric을 사용해야하는 것은 아니다. 즉, metric의 세 가지 성질을 만족하지 않는 measure일지라도 두 데이터가 얼마나 '비슷하냐'를 measure할 수 있는 measure라면 KNN에 적용할 수 있다.

K-means Clustering

군집화는 몇 개의 그룹으로 나누어야 할 지가 관건이다. 결정 방법으로 **엘보우와 실루엣** 기법 등이 있다.

방법1) 엘보우(elbow) 기법

k-means 클러스터링은 클러스터 내 오차제곱합(SSE)의 합이 최소가 되도록 클러스터의 중심을 결정해 나가는 방법이다. 만약 클러스터의 개수를 1로 두고 계산한 SSE 값과, 클러스터의 개수를 2로 두고 계산한 SSE 값을 비교했을 때, 클러스터의 개수를 2로 두고 계산한 SSE 값이 더 작다면 1개의 클러스터 보다 2개의 클러스터가 더 적합하다고 볼 수 있다. 이런 식으로 클러스터의 개수를 늘려가면서 계산한 SSE를 그래프로 그려보면 SSE의 값이 점점 줄어들다가 어느 순간 줄어드는 비율이 급격하게 작아지는 부분이 생기는데, 그래프의 모양을 보면 팔꿈치 꼬트머리 처럼 보이는 부분이 있는데 이 부분이 우리가 구하려는 최적의 클러스터 개수가 된다.



방법2) 실루엣(silhouette) 기법

클러스터링의 품질을 정량적으로 계산해 주는 방법이다. 클러스터의 개수가 최적화되어 있으면 실루엣 계수의 값은 1에 가까운 값이 된다. 실루엣 기법은 k-means 클러스터링 기법 이외에 다른 클러스터링에도 적용이 가능하다