1. 머신러닝 (Machine Learning)

말 그대로 '머신러닝' 이란 기계를 학습시키는 과정이다. 수많은 학습데이터를 통해서 기계를 학습시키고 이를 통해 원하는 결과를 얻는 것이다. 예를 들어 나무 그림을 생각해보자. 이 그림을 나무라고 인식하기 위해선, 컴퓨터에게 수많은 나무데이터와 나무가 아닌 데이터를 학습시킨다. 이렇게 학습이 된 후, 새로운 그림을 주어줄 때 나무인지 아닌지 구별할 수 있게끔 하는 방식이다.

일반적으로 컴퓨터를 통해서 원하는 결과를 얻기 위해서는 일련의 알고리즘이 필요하다. 알고리즘이 미리 주어져 있으면 이를 이용해도 되지만, 대부분의 경우에는 미리 주어진 경우가 없기 때문에 알고리즘 설계의 필요성이 요구된다. 예를 들어, 소비자가 마트에서 어떤 물품을 살지 예측하는 것은 정확한 알고리즘이 존재하지 않는다. 여기서 중요한 것은 특정한 알고리즘은 알 수 없지만, 소비자들이 지금까지 마트에서 샀던 물품들에 관한 데이터는 존재한다는 것이다. 이러한 과거의 데이터를 학습시키고 쓸만한 알고리즘을 설계한다.

즉, 학습데이터를 가지고 어떤 모델을 정해준다면 컴퓨터는 스스로 그 모델에 가장 적합한 파라미터를 찾는다. 파라미터란 모델을 결정하는 변수라 생각하면 된다. 학습데이터를 사용하고 파라미터가 정해져 있지 않다는 것이 전통적인 인공지능과 머신러닝의 차이점이다. 전통적인 인공지능은 모델에 필요한 파라미터를 전문가가 직접 입력한다. 그렇기 때문에 상황에 따른 응용이나 융통성이 떨어진다. 이에 비해서 머신러닝은 학습을 통해 추론을 하는 것이고, 새로운 데이터가 들어왔을 때에도 파라미터를 수정해가면서 계속 적용할 수 있다. 결국, 지식을 심어주는 것이 아니라 컴퓨터가 지식을 찾아가는 것이다. 머신러닝기법을 빅데이터에 적용하는 것을 데이터마이닝이라고 한다.

예를 들어 최초 학습데이터에 소나무, 잣나무 의 데이터만 존재한다고 생각해보자. 전통적인 인공지능은 오직 소나무, 잣나무만 구별해 낼 수 있고 새로운 종류의 나무가 들어오면 '나무'라고 인식할 수 없다. 반면 머신러닝은 소나무, 잣나무의 특징을 통해 '나무'라는 대표 특성을 찾아내려고 하고, 이를 통해 새로운 종류의 나무(ex, 밤나무) 가 주어졌을 때에도 나무라고 구별할 수 있게 된다.

2. 머신러닝의 종류

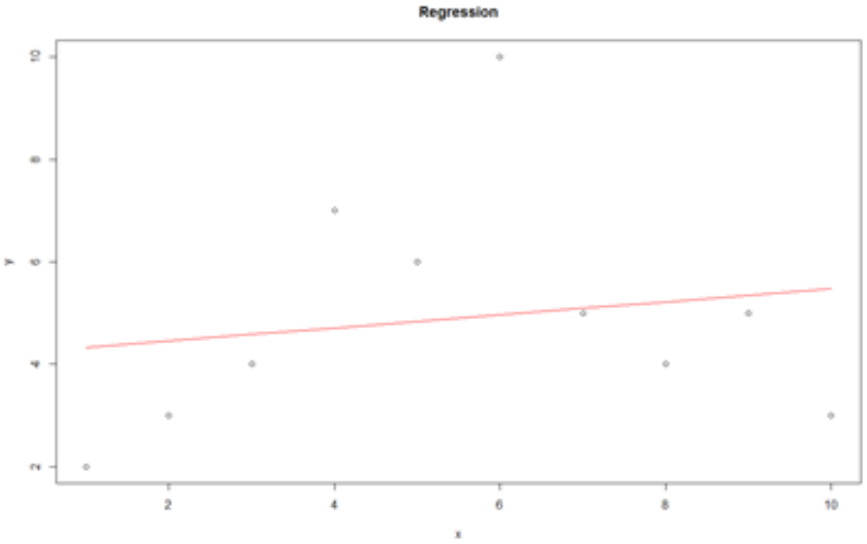
(a) 지도학습 (Supervised learning)

y 라는 output 이 정해져 있다. 학습데이터가 x 라는 input 과 y 라는 output 의 쌍으로 구성되어 있으며, x 가 주어졌을 때 y 라는 결과를 예측하는 것이 핵심이다. 학습데이터의 x 와 y 를 통해 모델을 만들고, 새로운 x 가 들어왔을 때 y 를 예측하는 것이 지도학습의 주요 목표라 할 수 있다.

.

Classification : 지도학습 중에서 output 이 범주형 변수인 경우, 학습데이터를 통해서 학습을 진행하고, 새로운 데이터 (x 라는 input) 가 어떤 범주(y 라는 output) 에 속하는지 예측한다.

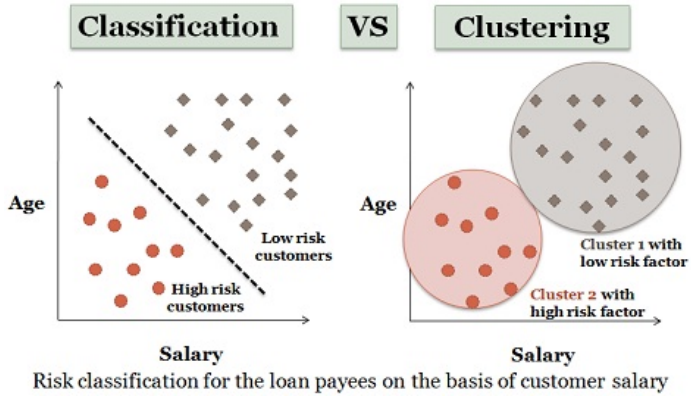
Regression : 지도학습 중에서 output 이 연속형 변수인 경우, 학습데이터를 통해서 학습을 진행하고, 새로운 데이터가 어떤 값을 가질지 예측한다. 한국말로 회귀분석이라고 불리는데 이는 아래와 같은 데이터가 있을 때, 적어도 2 차함수는 되어야 데이터를 설명할 수 있다. 이를 더 낮은 차원인 직선으로 설명하는 것이 선형회귀이다. 즉, 더 낮은 차원으로 회귀한다 혹은 퇴보한다는 의미에서 회귀분석이라는 이름의 의미를 이해할 수 있다. 결론적으로, 복잡한 데이터를 단순한 형태로 설명하는 것이다.



(b) **비지도학습 (Unsupervised learning)** : y 라는 output 이 없다. 학습데이터에 x 만 있다고 생각할 수 있고, 흥미로운 패턴을 발견하거나 집단을 나누는 등 서술적인 측면이 강하다.

Clustering : 학습데이터의 다양한 변수들을 기준으로 여러 개의 군집을 나누는 것이다. y 가 주어져 있지 않기 때문에, 각 군집 별 특성을 임의로 지정해주고, 이를 각 군집의 y 처럼 사용한다.

※ Classification vs Clustering



Risk classification for the loan payees on the basis of customer salary

위의 그림과 같이 Age 와 Salary 라는 변수를 기반으로 Classification 과 Clustering 을 진행하는데 Classification 은 이미 Low risk 와 High risk 라는 이름이 정해져 있다. 이를 통해 이 변수들이 두 집단을 적절하게 나누는가를 판단할 수 있다. 반면에 Clustering 은 변수들을 통해서 군집을 형성하고 이 군집들이 서로 어떤 특징을 가질까 해석하고 분석하는 것이다.

Association rule : 연관규칙분석이라고 불리며, 조건부확률에 기반하여 아이템간의 규칙을 발견하는 것이다. 예를 들어, 치킨을 시킬 때 맥주를 같이 시키는 경향이 있다는 식의 규칙이다. 장바구니분석이라고도 불린다. 대표적으로 지지도(support), 신뢰도(confidence), 향상도(lift) 와 같은 여러 방식을 사용하여, 상호간 연관성을 확인한다.

지지도(support) : $\Pr(A \cap B) = \frac{A와\ B\ 둘다\ 고른\ 사람\ 수}{전체\ 분석하는\ 총\ 인원\ 수}$

신뢰도(confidence) : $\Pr(A \mid B) = \frac{A와\ B\ 둘다\ 고른\ 사람\ 수}{B를\ 고른\ 사람\ 수}$

향상도(lift) $\frac{\Pr(A|B)}{\Pr(B)} = \frac{\Pr(A \cap B)}{\Pr(A) * \Pr(B)}$

각 방식에 임의의 기준을 정하고, 이를 기반으로 적당한 연관성을 찾는 방식을 일컫는다. 예를 들어 지지도는 0.6 이상, 신뢰도는 0.7 이상 중에서 향상도가 제일 높은 것은 뭘까? 와 같은 방식이라 할 수 있다.