# 기계학습 문제 분류와 알고리즘

## 기계학습 문제 분류

주어진 문제를 해결하는 수 있도록 수집한 데이터를 수학적 모델을 통해 기계를 학습시키는 것을 기계학습이라 한다. 기계학습 알고리즘은 크게 지도학습, 비지도학습, 강화학습으로 나누어지며, 많이 사용하는 지도학습과 비지도학습에 대해서 알아보도록 한다.

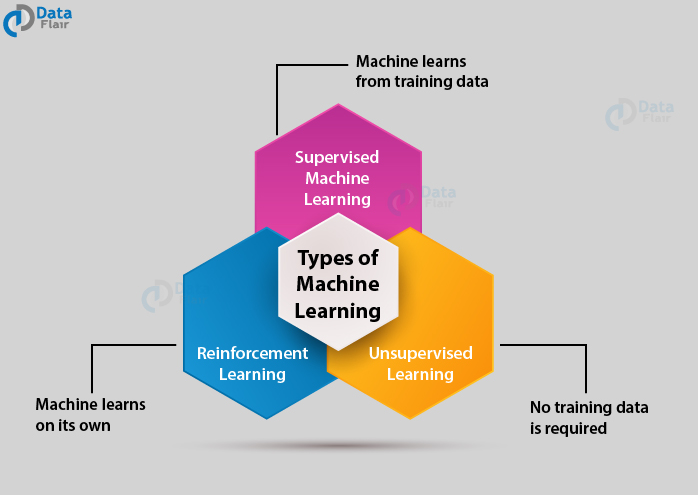


그림 1. 기계학습 문제 분류

### 지도학습

실용적인 기계학습의 대부분은 지도학습을 사용한다. 지도학습은 수학적 모델(지도학습 알고리즘)을 가지고 학습데이터 내의 정답과 입력 간의 관계를 파악하여, 주어진 문제에 대한 올바른 정답을 예측하는 방법이다. 따라서, 지도학습에 쓰이는 학습데이터는 정확한 정답과 문제에 대한 힌트를 가져야 한다. 지도학습은 주로 분류, 회귀문제의 유형이 적용된다.

### 비지도학습

비지도학습은 관측치들의 특성 정보를 담고 있는 학습데이터를 주로 사용하며, 지도학습과는 달리 학습데이터에 정답이 없다. 비지도학습은 학습데이터의 구조, 분포, 패턴, 특성 등을 파악한다. 주로 어떤 대상들을 구분해서 그룹을 만드는 군집화 문제, 서로 연관된 특징을 찾아내는 연관 문제 즉, 탐색적인 목적의 문제의 유형에 적용된다.

## 지도학습 알고리즘

지도학습의 다양한 알고리즘 중 서포트 벡터 머신과 인공신경망 알고리즘에 대해 알아보겠다. 두 기법 모두 간단한 구조에서부터 시작할 수 있고 분류와 회귀 문제에서 뛰어난 성능을 보인다. Part3에서 서포트 벡터 머신(Chap3-1,2)과 인공신경망(Chap2-1,2, Chap4-1,2)이 화학공학에서 적용되는 예제를 확인할 수 있다.

### 서포트 벡터 머신(Support vector machine, SVM)

서포트 벡터 머신은 분류 문제에 주로 적용되는 알고리즘이다. 학습이 진행되면서 학습데이터를 클래스로 분리하는 초평면(hyperplane)을 최적으로 분리하는 선형분리를 찾는다. 두 클래스 사이에 서로 가장 가까운 점을 서포트 벡터라고 한다. 서포트 벡터와 초평면과의 거리를 마진이라고 부르며, SVM의 목표는 이 마진을 최대화시키는 것이다. 즉, 마진이 최대에 도달하면 그때의 초평면이 최적의 값을 갖는다.

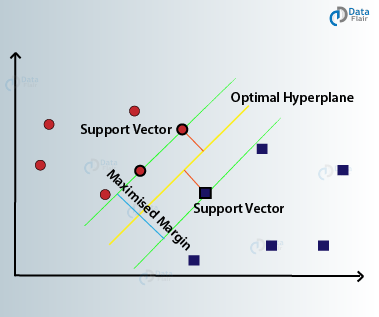


그림 . 선형 SVM

SVM 알고리즘은 학습 데이터에서의 서포트 벡터가 클래스 사이의 결정 경계(decision boundary)를 구분하는데 얼마나 중요한지를 학습시켜준다. 위의 그림의 경우에는 데이터가 2차원이지만 초평면은 1차원이다. 더 높은 차원의 학습데이터, 예를 들어 n차원의 데이터라면 이를 구분하는 초평면은 n-1차원의 값을 가질 것이다.

### 인공신경망(Artificial neural network, ANN)

인공신경망은 신경계의 뉴런이 과거 데이터에서 학습하는 것처럼 인간의 두뇌를 모델로 한 특수한 유형의 지도학습 알고리즘이다. 인공신경망은 입력과 출력 사이의 관계를 알아내기 위한 비선형 모델이며, 이미지 인식, 음성 인식, 기계 번역 등 다양한 곳에 적용될 수 있다.

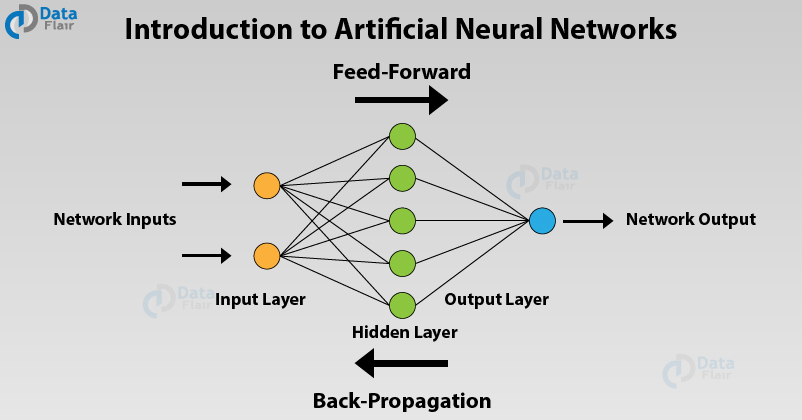


그림 . ANN의 구조

그림 3과 같이 인공신경망은 입력, 은닉, 출력 총 3개의 층을 가진다. 각 층은 노드들로 구성되어 있으며, 어떤 값을 입력 받고 출력하는 역할을 한다. 입력 층은 이미지 픽셀, 텍스트, 오디오 파일 등의 형태의 정보를 받게 되는 층을 말한다. 은닉 층은 하나 혹은 여러 층으로 구성될 수 있으며, 입력 층에 들어온 데이터와의 수학적인 계산을 수행하고 패턴을 인식한다. 마지막으로 출력 층은 앞선 층에서 수행된 계산을 통해 얻은 결과를 출력한다. 출력 층 노드의 수는 문제의 종류에 따라서 결정된다.

인공신경망에서 각 노드에는 할당된 가중치와 편향이 있다. 입력 층에서의 입력 노드는 입력된 값을 그대로 출력하고, 그 값을 은닉 노드에서 받아 특정 함수로 변환하여 출력하게 된다. 이 함수를 활성화 함수라고 한다. 활성화 함수는 보통 은닉과 출력 노드에 존재하며, 해당 노드에 입력된 값을 변환하여 출력하는 역할을 한다. 출력되는 값을 해당 노드가 정답을 맞추는데 기여하는 정도를 반영한다. 주로 많이 쓰이는 함수로는 Sigmoid, RELU, Softmax, tanh 등이 있다.

신경망 알고리즘은 Feedforward 신경망과 Feedback 신경망 두 가지 유형이 있다. Feedforward 인공신경망은 정보의 흐름이 입력 층에서 은닉 층, 그리고 출력 층으로 한 방향으로만 발생한다. 이 유형의 신경망은 분류, 이미지 인식과 같은 지도 학습에 주로 사용된다. 데이터가 순차적이지 않은 경우에 사용된다. Feedback 인공신경망은 출력 층의 출력이 다시 입력 층으로 입력되는 feedback이 작용되며, Recurrent neural network의 경우와 같이 기억 유지를 한다. 이 유형은 데이터가 순차적이거나 시간 종속적인 영역에 주로 사용된다.

## 비지도학습 알고리즘

비지도학습의 다양한 알고리즘 중 주성분 분석 기법과 오토인코더 알고리즘에 대해 알아보겠다. 비지도 학습은 고차원의 데이터를 저차원으로 축소시키는 대표적인 차원축소 기법이고, 오토인코더는 비선형 매니폴드를 학습하여 차원축소 뿐만 아니라 노이즈 제거 등에도 활용될 수 있다. Part3에서 주성분 분석(Chap2-3)과 오토인코더(Chap5-1)가 화학공학에서 적용되는 예제를 확인할 수 있다.

### 주성분 분석(Principle components analysis, PCA)

주성분 분석은 비지도 기능 학습 알고리즘의 속도를 크게 높이는 데 사용할 수 있는 차원 축소 알고리즘이다. 각각의 축이 주요 요소들을 표현하는 n-차원의 타원면에 데이터를 맞추는 것으로 생각할 수 있다. 여기서 우리의 데이터 집합의 표현에서 몇 개의 작은 축과 그에 해당하는 주요 요소들을 생략함으로써 우리는 상응하는 작은 정보만을 잃을 수 있다. 더욱이, 주성분 분석을 통해 많은 알고리즘의 중요한 사전 처리 단계를 구현할 수 있다.

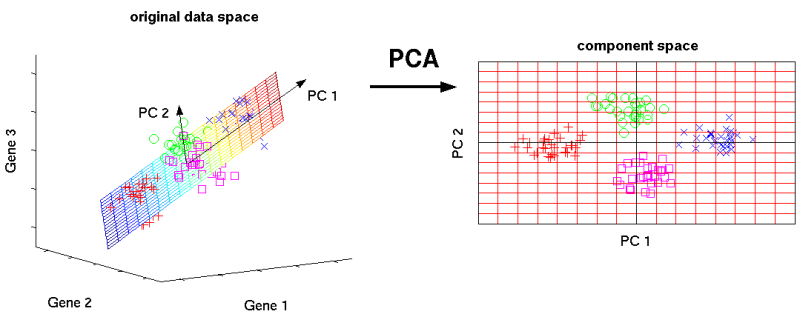


그림 . 주성분 선택 방법과 시각화

이미지에 대한 알고리즘을 훈련한다고 가정하면 이미지의 인접 픽셀 값이 높은 상관 관계를 갖기 때문에 입력이 다소 중복된다. 구체적으로, 우리가 16x16 회색조 이미지 패치에 대해 훈련한다고 가정할 때, x는 256차원 벡터이며, 하나의 특징 xj는 각 픽셀의 강도에 해당한다. 인접 픽셀 간의 상관 관계로 인해 PCA를 사용하면 오류가 거의 발생하지 않으면서도 훨씬 낮은 차원으로 입력을 근사화 할 수 있다.

### 오토인코더(Autoencoder)

오토인코더 신경망은 역전파를 적용하여 출력값을 입력값과 동일하게 설정하는 비지도 학습 알고리즘이다.

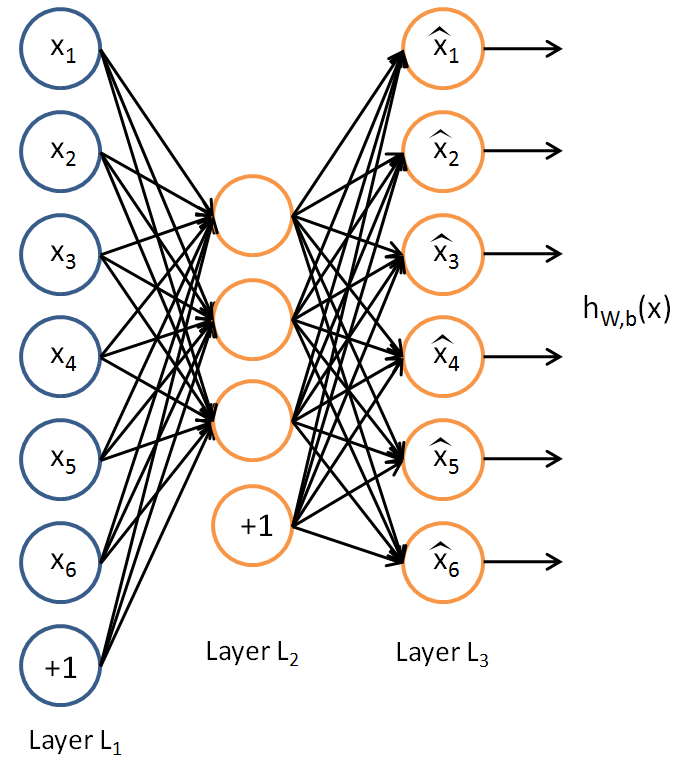


그림 . 오토인코더 구조

그림 5와 같이 오토인코더는 함수 hw,b(x)≈x를 학습하려고 한다. 즉, 항등 함수에 대한 근사를 학습하여 x와 유사한 x^를 출력하려고 한다. 그러나 은닉 유닛의 수를 제한하는 것과 같이 네트워크에 제약을 가함으로써 데이터에 대한 흥미로운 구조를 발견할 수 있다. 구체적인 예로서, 입력 x가 10×10 이미지(100픽셀)의 픽셀 강도 값이므로 n=100이고 레이어 L2에 50개의 은닉 유닛이 있다고 가정하면, 은닉 유닛이 50개뿐이므로 네트워크는 입력의 "압축된" 표현을 학습해야 한다. 즉, 은닉 유닛 활성화 벡터만 주어지면 100픽셀 입력 x를 "'재구성''할 수 있어야 한다. 입력이 완전히 무작위인 경우에는 이 압축 작업이 매우 어렵겠지만 입력 기능 중 일부에 상관 관계가 있는 경우에는 오토인코더가 이러한 상관 관계를 학습할 수 있다. 사실, 간단한 오코인코더는 종종 PCA와 매우 유사한 저차원 표현을 학습하게 된다.

위에서는 은닉 유닛의 수가 작은 경우를 다루었지만, 은닉 유닛의 수가 많을 때에는 네트워크에 다른 제약 조건을 부과함으로써 흥미로운 구조를 발견할 수 있다. 특히 은닉 유닛에 "희소성" 제약 조건을 부과하면 오토인코더는 은닉 유닛의 수가 많더라도 데이터에서 여전히 흥미로운 구조를 발견할 수 있다.

## 알고리즘 선택

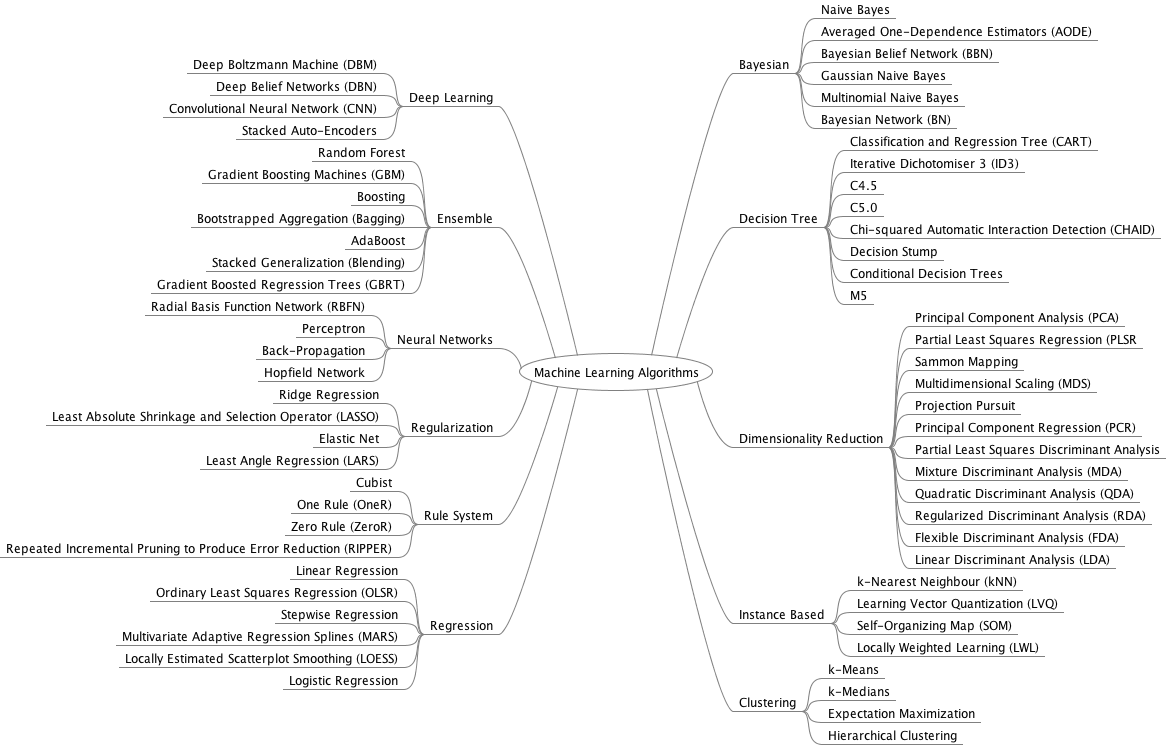


그림 6. 다양한 머신러닝 알고리즘

그림 6과 같이 무수히 많은 알고리즘이 개발되어 있고 현재도 끊임없이 발전 중이다. 대중적인 알고리즘들은 배포된 라이브러리를 설치하여 사용할 수 있고 최신 알고리즘들은 github 등에 공개되어 있는 오픈소스 코드를 활용하여 구현할 수 있다.

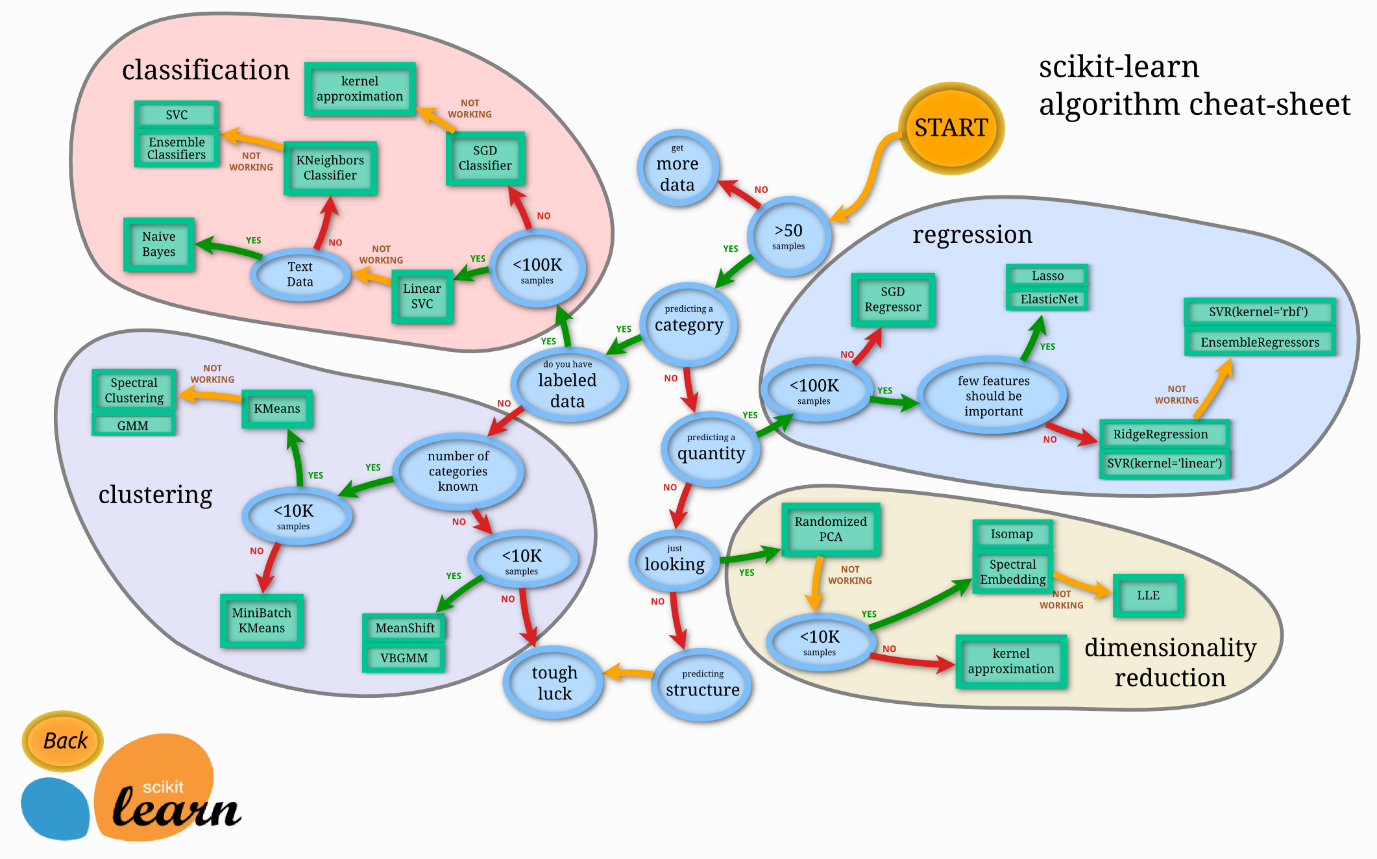


그림 7. 머신러닝 알고리즘 선택 방법

무수히 많은 알고리즘들 중에 어떤 알고리즘을 사용할 것인가는 문제의 세부 조건에 따라 그림 7을 참고하여 결정할 수 있다. 데이터의 종류, 크기, 속성, 라벨 여부 등이 알고리즘 선정에 영향을 미치게 된다.

## 결론

본 장에서는 기계학습 문제를 분류하고 분류에 따른 알고리즘들을 살펴보았다. 풀고자 하는 문제를 정확히 파악하고 분류한 후에 어떤 알고리즘이 문제에 적합한지 판단할 수 있다. 다양한 머신러닝 알고리즘이 있지만 지도학습의 서포트 벡터 머신과 인공신경망, 비지도학습의 주성분 분석과 오토인코더에 대해 간략하게 알아보았다.