# 기계학습 문제 분류와 알고리즘

## 기계학습 문제 분류

기계학습이란 명시적인 컴퓨터 프로그램 및 알고리즘 개발로 수행하기 어려운 복잡계 문제를 해결하기 위하여 사용되는 방법론이다. 특히, 소재 특성 분석, 공정 개발, 제어, 최적화 등 다수의 파라미터로 구성된 화공 산업의 다양한 의사결정 문제를 해결하는데 매우 적합한 방법론이다. 그림 1과 같이 기계학습 알고리즘은 학습을 시키는 방식에 따라 크게 지도학습, 비지도학습, 강화학습으로 나뉜다. 본 절에서는 지도학습과 비지도학습의 주요 개념과 특징에 관해 살펴본다.

텍스트, 화이트보드이(가) 표시된 사진

자동 생성된 설명

그림 1. 기계학습 문제 분류

### 지도학습

화공 산업의 다양한 문제 중, 상대적으로 많은 데이터가 확보 가능하고 데이터의 특징(feature)과데이터 간의 관계가 높은 정합성을 보이는 문제의 경우, 주로 지도학습을 사용하여 문제 해결이 가능하다. 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 크게 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 등 탐색적인 목적의 문제 유형에 적용된다.

## 지도학습 알고리즘

지도학습 방법론에는 매우 다양한 알고리즘이 적용되고 있으며, 다양한 분야에서 해당 분야에 적합한 새로운 기능이 탑재된 새로운 알고리즘 개발이 꾸준히 개발되고 있다. 대표적인 지도학습 알고리즘으로 서포트 벡터 머신(support vector machine; SVM)과 인공신경망(artificial neural network; ANN) 알고리즘이 있다. 두 기법 모두 구조가 간단함에도 불구하고, 분류와 회귀 문제 해결에서 매우 뛰어난 성능을 보인다. 상기 두 기법을 이용한 실제 화공 산업 문제 해결의 예는 3 장 서포트 벡터 머신(3-1장, 3-2장)과 2 장 및 4 장 인공신경망(2-1장, 2-2장, 4-1장, 4-2장)에서 학습할 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 선형분리 해를 찾는 과정이 알고리즘의 핵심이다. 즉, 구성된 다양한 클래스 사이에 서로 가장 가까운 점들인 서포트 벡터과 초평면과의 거리(마진)를 최대화하기 위하여 사용 데이터로 구성된 차원을 분석한다. 서포트 벡터 머신 알고리즘은 마진이 최대화되는 초평면을 계산함으로써 비정형의 데이터를 분류 및 설명하게 된다.

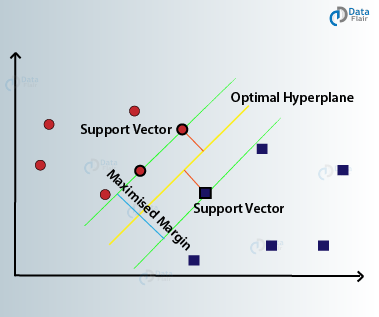


그림 2. 선형 서포트 벡터 머신

서포트 벡터 머신 알고리즘은 학습 데이터에서의 서포트 벡터가 클래스 사이의 결정 경계(decision boundary)를 구분하는데 얼마나 중요한지를 학습시켜준다. 위 그림의 경우, 데이터가 2 차원이지만 초평면은 1 차원이다. 더 높은 차원의 학습데이터, 예를 들어 n차원의 데이터라면 이를 구분하는 초평면은 n-1 차원의 값을 가지게 된다.

### 인공신경망(artificial neural network, ANN)

인공신경망은 여러 층으로 구성된 인간의 두뇌에 착안하여 개발된 특수한 유형의 지도학습 알고리즘이다. 인공신경망은 비선형 모델로서, 이미지 인식, 음성 인식, 기계 번역 등 다양한 곳에 적용될 수 있다.

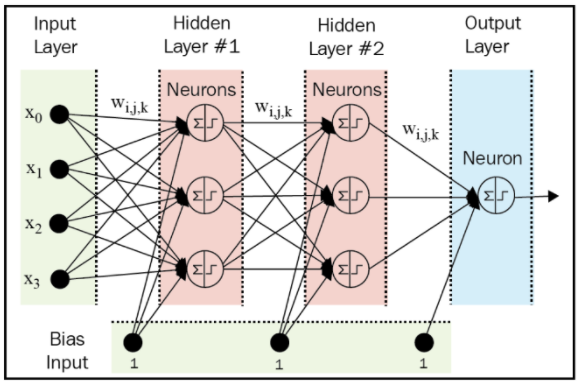


그림 3. 인공신경망의 구조

그림 3과 같이 인공신경망은 입력(input), 1 개 이상의 은닉(hidden), 출력(output) 등 총 3 개 이상의 층(layer)을 가진다. 각 층은 다수의 노드(perceptron)들로 구성되어 있으며, 입력된 데이터의 특징을 분석하여 다음 층의 입력 데이터로 출력하는 역할을 한다. 입력 층은 이미지 픽셀, 텍스트, 오디오 파일 등의 형태 정보를 받게 되는 층을 말한다. 은닉 층은 하나 혹은 여러 층으로 구성될 수 있으며, 입력 층에 입력된 데이터의 수학적인 계산을 수행하고 패턴을 인식한다. 마지막으로 출력 층은 앞선 층에서 수행된 계산을 통해 얻은 결과를 출력한다. 출력 층 노드의 수는 문제의 종류에 따라 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 함수(activate function)라고 한다. 활성화 함수는 입력 값으로부터 다음 계층으로 전달하는 출력신호를 생성하는 함수로, 보통 은닉과 출력 노드에 존재한다. 주로 많이 쓰이는 함수로는 Sigmoid, RELU, Softmax, tanh 등이 있다. 신경망 알고리즘은 데이터 처리 순서에 따라 크게 Feedforward 신경망과 Feedback 신경망 두 가지 유형으로 나뉜다. Feedforward 인공신경망은 정보의 흐름이 입력 층에서 은닉 층, 그리고 출력 층으로 한 방향으로만 발생한다. 이 유형의 신경망은 분류, 이미지 인식과 같이 데이터가 순차적이지 않은 지도 학습에 주로 사용된다. Feedback 인공신경망은 출력 층의 출력이 다시 입력 층으로 재순환되는 신호의 되먹임이 발생하며, Recurrent neural network의 경우와 같이 기억 유지를 한다. 이 유형은 데이터가 순차적이거나 시간 종속적인 영역에 주로 사용된다.

## 비지도학습 알고리즘

비지도학습의 대표적인 알고리즘으로 주성분 분석(principle component analysis; PCA)과 오토인코더(auto encoder)가 있다. 주성분 분석은 고차원의 데이터를 저차원으로 축소시키는 대표적인 차원 축소 기법이고, 오토인코더는 비선형 매니폴드를 학습하여 차원축소와 함께 노이즈 제거에도 효과적이다. 비지도학습 알고리즘을 통한 실제 사례는 2 장의 주성분 분석(2-3장)과 5 장의 오토인코더(5-1장)에서 학습 가능하다.

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

주성분 분석은 차원 축소 기법으로서 비지도학습의 속도를 개선하는데 사용될 수 있다. 주성분 분석은 전체 데이터를 n 차원의 주요 요소들로 표현하는 기법으로, 정보의 작은 손실만으로 데이터의 간략화가 가능하며, 많은 알고리즘의 적용과정에서 사전처리단계로 활용될 수 있다.

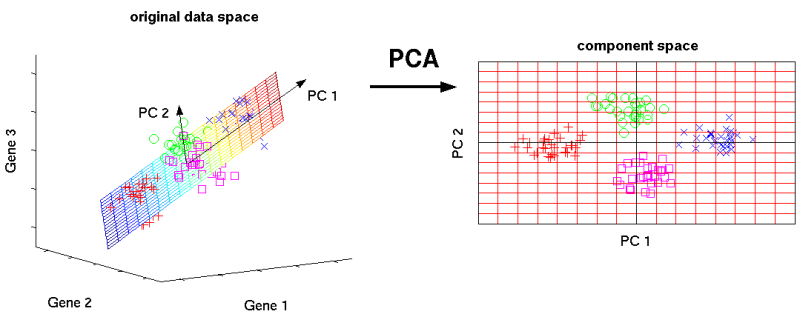


그림 4. 주성분 선택 방법과 시각화

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 256 차원 벡터이며, 하나의 특징 xj는 각 픽셀의 강도에 해당한다. 인접 픽셀 간의 상관 관계로 인해 주성분 분석을 사용하면 오류를 최소화하면서도 훨씬 낮은 차원으로 입력을 근사화 할 수 있다.

### 오토인코더(Autoencoder)

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

그림 5 와 같이 오토인코더는 함수 hw,b(x)≈x 를 학습한다. 즉, 항등함수에 대한 근사를 학습하여 x 와 유사한  를 출력한다. 그러나 은닉 유닛의 수를 제한하는 것과 같이 네트워크에 제약을 가함으로써 데이터에 대한 흥미로운 구조를 발견할 수 있다. 한 예로, 입력 x 가 10×10 이미지(100픽셀)의 픽셀 강도 값이므로 n=100 이고 레이어 L2 에 50 개의 은닉 유닛이 있다고 가정하면, 은닉 유닛이 50 개뿐이므로 네트워크는 입력의 "압축된" 표현을 학습해야 한다. 즉, 은닉 유닛 활성화 벡터만 주어지면 100픽셀 입력 x를 "'재구성''할 수 있어야 한다. 입력이 무작위인 경우에는 이 압축 작업이 매우 어렵겠지만, 입력 기능 중 일부에 상관 관계가 있는 경우에는 오토인코더가 이러한 상관 관계를 학습할 수 있다. 간단한 오코인코더는 종종 주성분 분석과 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.

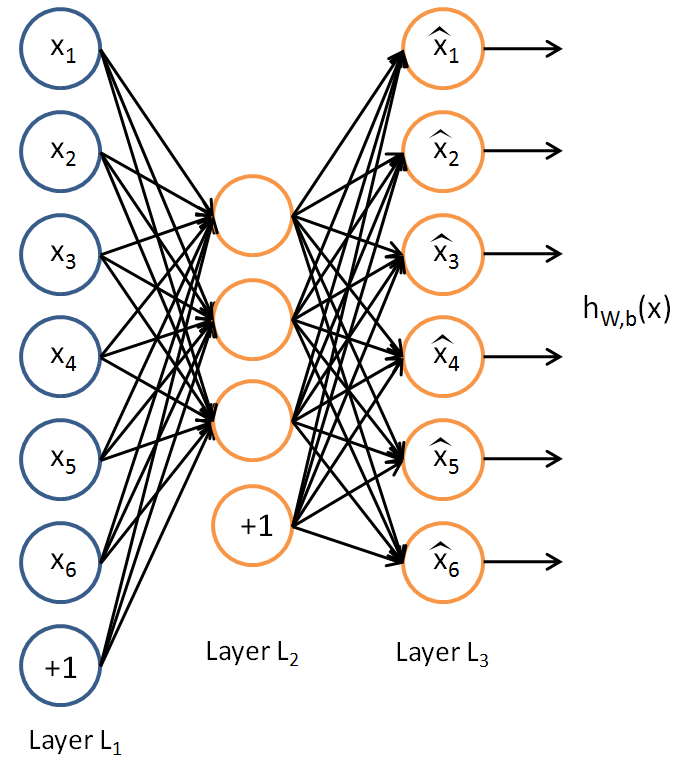


그림 5. 오토인코더 구조

## 알고리즘 선택

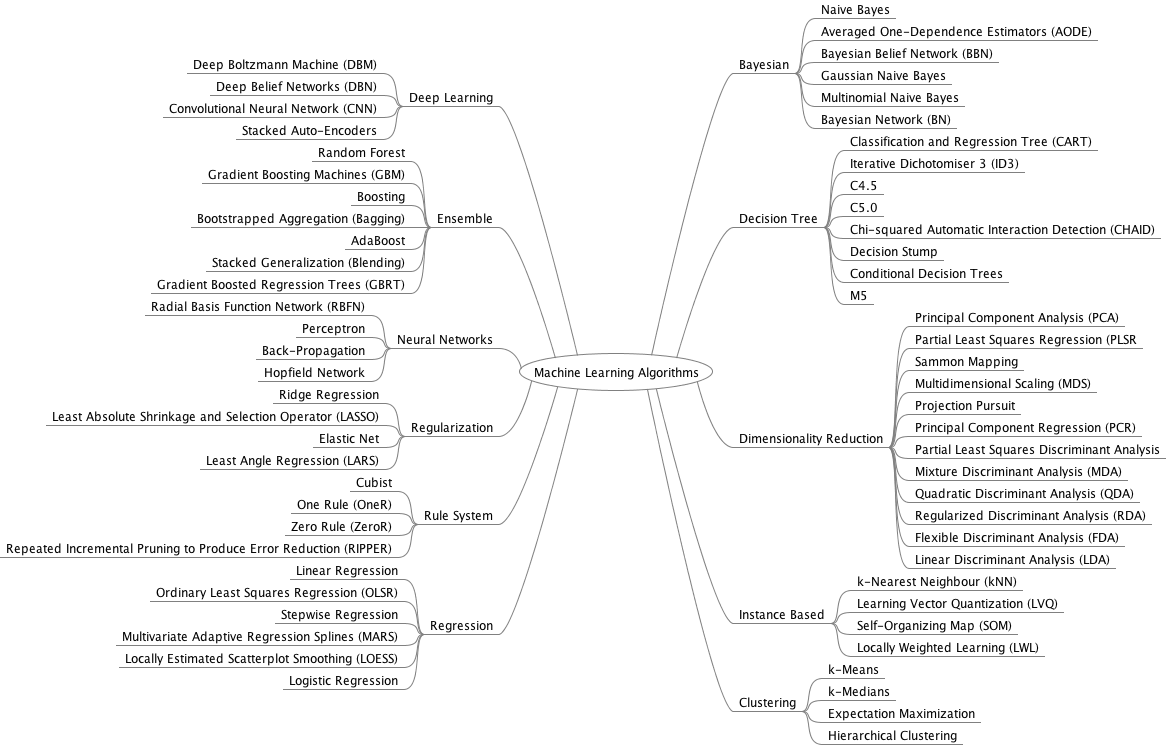


그림 6. 다양한 머신러닝 알고리즘

그림 6 과 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.

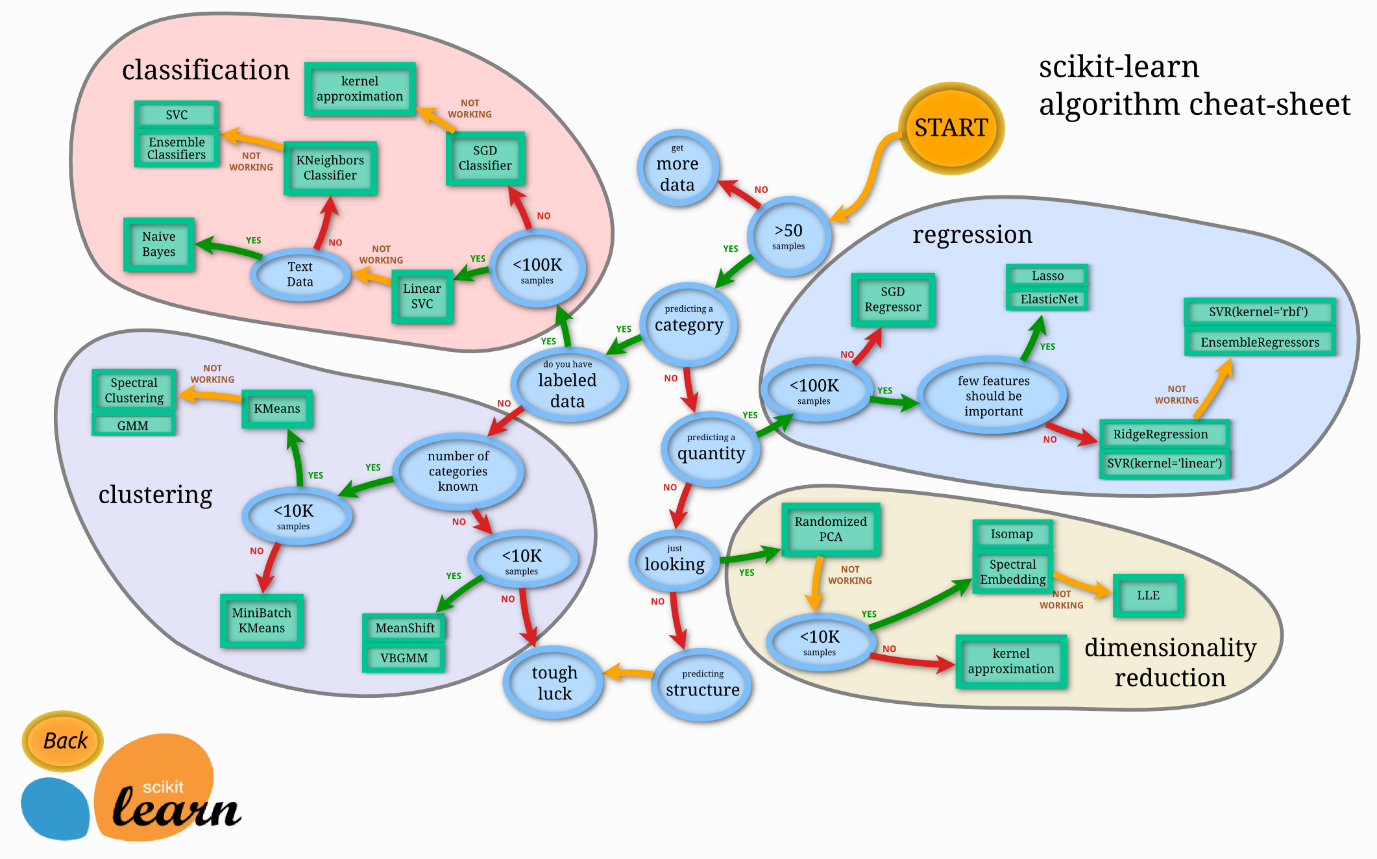


그림 7. 머신러닝 알고리즘 선택 방법

무수히 많은 알고리즘들 중에 어떤 알고리즘을 사용할 것인가는 문제의 세부 조건에 따라 그림 7 을 참고하여 결정할 수 있다. 데이터의 종류, 크기, 속성, 라벨 여부 등이 알고리즘 선정에 영향을 미치게 된다. 먼저 얻고자 하는 출력값이 카테고리인지 수치인지에 따라 크게 나누어질 수 있다. 각각에서는 정답(라벨)의 유무에 따라 지도 학습 알고리즘을 사용할지 비지도 학습 알고리즘을 사용할지 달라진다. 모델의 목적이 카테고리를 나누는 것이라면 라벨이 존재할 때는 분류 알고리즘을 사용할 수 있고 라벨이 존재하지 않는다면 클러스터링 알고리즘을 사용할 수 있다. 마찬가지로 수치형 데이터를 사용할 때도 학습할 때 사용할 라벨이 존재하고 예측 모델을 만들고 싶다면 회귀 알고리즘을 사용할 수 있고 라벨이 존재하지 않지만 시각화 등을 통해 표현하고 싶다면 차원 축소 알고리즘을 선택할 수 있다. 데이터의 종류나 라벨 여부는 알고리즘을 선택하는데 가장 기본적으로 고려하는 사항이고 그 외에 일반적인 규칙이 존재하는 것은 아니기 때문에 다양한 알고리즘을 사용해보며 데이터셋의 특징을 적절히 반영해줄 수 있는 알고리즘을 찾아야한다.

## 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.

## 학습 결과

* 학습 내용

대표적인 기계학습 알고리즘의 원리와 특징 익히기

* 학습 결과 확인하기

문제 특징에 따라 적절한 알고리즘 선택하기

* 학습 결과 응용하기

각 알고리즘의 최신 동향 파악하기