# 기계학습

## 정의

* 경험을 통해서 나중에 유사하거나 같은 일을 더 효율적으로 처리 할 수 있도록 시스템의 구조나 파라미터를 바꾸는 것
* 컴퓨터가 데이터로부터 특정 문제해결을 위한 지식을 자동으로 추출해서 사용할 수 있게 하는 기술
* 연역적 학습 : 연역적 추론을 통한 학습
* 귀납적 학습 : 사례들을 일반화하여 패턴 또는 모델을 추출하는 것

## 종류

### 지도 학습

* + 입력-출력의 데이터들로부터 새로운 입력에 대한 출력을 결정할 수 있는 패턴 추출
  + 분류 : 출력이 정해진 부류중의 하나로 결정
  + 회귀 분석 : 출력이 연속인 영역의 값 결정
  + 분류, 회귀, 군집화, 밀도추정, 차원 축소 문제

### 비지도 학습

* + 입력-출력의 데이터들로부터 새로운 입력에 대한 출력을 결정할 수 있는 패턴 추출
  + 분류 : 출력이 정해진 부류중의 하나로 결정
  + 회귀 분석 : 출력이 연속인 영역의 값 결정
  + 분류, 회귀, 군집화, 밀도추정, 차원 축소 문제

### 반지도 학습

* + 입력-출력의 데이터들로부터 새로운 입력에 대한 출력을 결정할 수 있는 패턴 추출
  + 분류 : 출력이 정해진 부류중의 하나로 결정
  + 회귀 분석 : 출력이 연속인 영역의 값 결정
  + 분류, 회귀, 군집화, 밀도추정, 차원 축소 문제

### 강화 학습

* + 출력에 대한 정확한 정보를 제공하지는 않지만, 평가 정보는 주어지는 문제에 대해 각 상태에서의 행동을 결정

## 대상 문제

### 분류

* + 데이터들을 정해진 몇개의 부류로 대응시키는 문제
  + 학습 데이터를 잘 분류할 수 있는 함수를 찾도록 학습한다.
  + 함수의 형태는 수학적이거나 규칙을 수 있다.

### 회귀

* + 학습데이터에 부합되는 출력값이 실수인 함수를 찾는 문제
  + 성능은 모델의 종류(함수의 종류)의 영향을 받음
  + 과적합 : 모델의 복잡도를 성능평가에 반영(목적 함수 = 오차의 합 + 가중치\*모델 복잡도

### 군집화

* + 유사성에 따라 데이터를 분할하는 것
  + 일단 군집화 : 데이터는 하나의 군집에만 소속
  + 퍼지 군집화 : 데이터는 여러 군집에 부분적으로 속하며 소속정도의 합은 1이 됨
  + 내제된 구조 추정, 전반적 구조 통찰
  + 가설 절정, 이상치 감지

### 밀도 추정

* + 부류별 데이터를 만들어 냈을 것으로 추정되는 확률 분포를 찾는 것
  + 각 부류별로 주어진 데이터를 발생시키는 확률 계산
  + 가장 확률이 높은 부류로 분류
  + 모수적 밀도 추정 : 분포가 특정 수학적 형태를 가지고 있다고 가정하여 파라미터 결정
  + 비모수적 밀도추정 : 분포에 대한 특정 함수를 가정하지 않고 주어진 데이터를 사용하여 밀도 함수의 형태 표현

### 차원 축소

* + 고차원의 데이터를 정보의 손실을 최소하하면서 저차원으로 변환하는 것
  + 차원의 저주 문제 완화 : 차원이 커질수록 거리분포가 일정해지는 경향 완화
  + 주성분 분석 : 분산이 큰 소수의 축들을 기준으로 데이터를 사상하여 저 차원으로 변환

### 이상치 탐지

* + 이상치 : 다른 데이터와 크게 달라서 다른 메커니즘에 의해 생성된 것이 아닌지 의심스러운 데이터
  + 잡음 : 관측오류, 시스템에서 발생하는 무작위적인 오차
  + 이상치 탐지 = 신규성 탐지와 관련
  + 점 이상치 : 다른데이터와 비교하여 차이가 큰 데이터
  + 상황적 이상치 : 상황과 맞지 않는 데이터
  + 집단적 이상치 : 여러데이터를 모아보면 비정상으로 보이는 데이터들의 집단
  + 부정사용감지 시스템 : 이상 거래 승인 요청시 카드 소유자에게 자동으로 경고 메세지 전송
  + 침입탐지 시스템 : 네트워크 트래픽을 관찰하여 이상 접근 식별