# 기계학습 알고리즘

## 결정트리

* 트리 형태로 의사결정 지식을 표현한 것
* 내부 노드 : 비교속성
* 간선 : 속성 값
* 단말 노드 : 부류, 대표 값

### 알고리즘

* + 단말 모든 데이터를 포함한 하나의 노드로 구성된 트리에서 시작
  + 반복적인 노드 분할 과정

### 분할 속성 선택

* + 분할한 결과가 가능하면 동질적인 것으로 만드는 속성 선택
  + 엔트로피 : 원래 정보량 측정 목적의 척도로 섞인 정도가 클수록 큰 값을 가짐.
  + 정보이득 : 정보 이득이 클 수록 우수한 분할 속성
  + 정보 이득 비 : 정보이득을 개선한 것으로 속성 값이 많은 속성에 대한 불이익
  + 지니 지수 : 데이터 집합에 대한 지니 값, 속성 A에 대한 지니 지수값 가중 평균

### 회귀

* + 단말노드가 부류가 아닌 수치값
  + 단말노드의 값은 해당 조건을 만족하는 것들이 가지는 대표값
  + 분할 속성 선택 : 표준편차 축소를 최대로 하는 속성 선택

## 앙상블 분류기

* 주어진 학습 데이터 집합에 대해서 여러 개의 서로 다른 분류기를 만들고, 이들 분류기의 판정 결과를 투표방식이나 가중치 투표방식으로 결합

### 배깅

* + 붓 스트랩을 통해 여러개의 학습데이터 집합을 만들고, 각 학습데이터 집합별로 분류기를 만드는 기법
  + 랜덤 포리스트 알고리즘 : 분류기로 다양한 결정트리를 사용하는 배깅 기법

### 부스팅

* + k개의 분류기를 순차적으로 만들어가는 앙상블 분류기 생성방법
  + 에이다 부스트

## K-근접 이웃 알고리즘

* (입력, 결과)가 있는 데이터들이 주어진 상황에서, 새로운 입력에 대한 결과를 추정할때 결과를 아는 최근접 k개의 데이터 결과정보를 이용한는 방법
* 분류 정확도에 따라 학습 데이터에 가중치를 변경해가면서 분류기 생성
* 학습 단계에서는 실질적인 학습이 일어나지 않고 데이터만 저장
* 새로운 데이터가 주어지면 저장된 데이터를 이용하여 학습(시간이 많이 걸림)

## 군집화 알고리즘

* 데이터를 유사한 것들끼리 모으는 것
* 군집 내의 유사도는 크게, 군집 간의 유사도는 작게
* 계층적 군집화 : 군집화의 결과가 군집들이 계층적인 구조를 갖도록 한는 것
* 분할 군집화 : 계층적 구조를 만들지 않고 전체 데이터를 유사한 것들끼리 나누어서 묶는 것(k-means 알고리즘)

## 단순 베이즈 분류기

* 부류 결정지식을 조건부 확률로 결정
* 가능도의 조건부 독립 가정