캐글을 통해 배우는 데이터 분석 개론

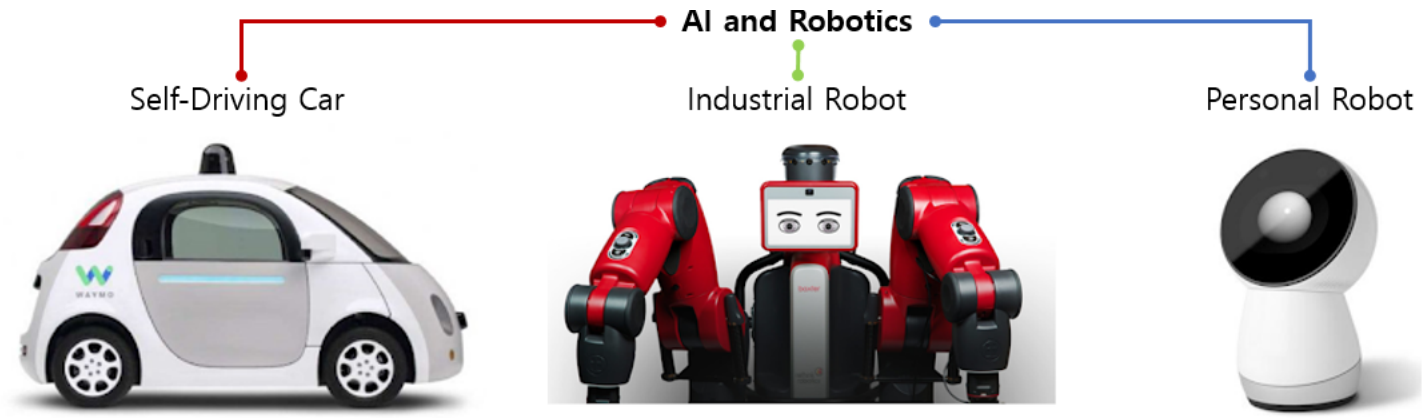
KISTI-세종대 빅데이터/AI 겨울 학교

연사 소개

- 이름: 최유경 (ykchoi@sejong.ac.kr)

- 학교: 세종대학교
- 학과: 지능기전공학부
  - 인공지능 기반의 기계/전자분야 융합형 인재양성을 목표로 함

- 연구실: 로보틱스 및 컴퓨터비전 연구실
  - http://www.rcv.sejong.ac.kr



강의 내용

- 파트1
  - 기계학습 이론 맛보기
  - 기계학습 실습 맛보기
  - 기계학습 경진대회 플랫폼(캐글) 맛보기

- 파트2
  - 기계학습 경진대회 플랫폼(캐글) 도전하기 I
  - 기계학습 경진대회 플랫폼(캐글) 도전하기 II

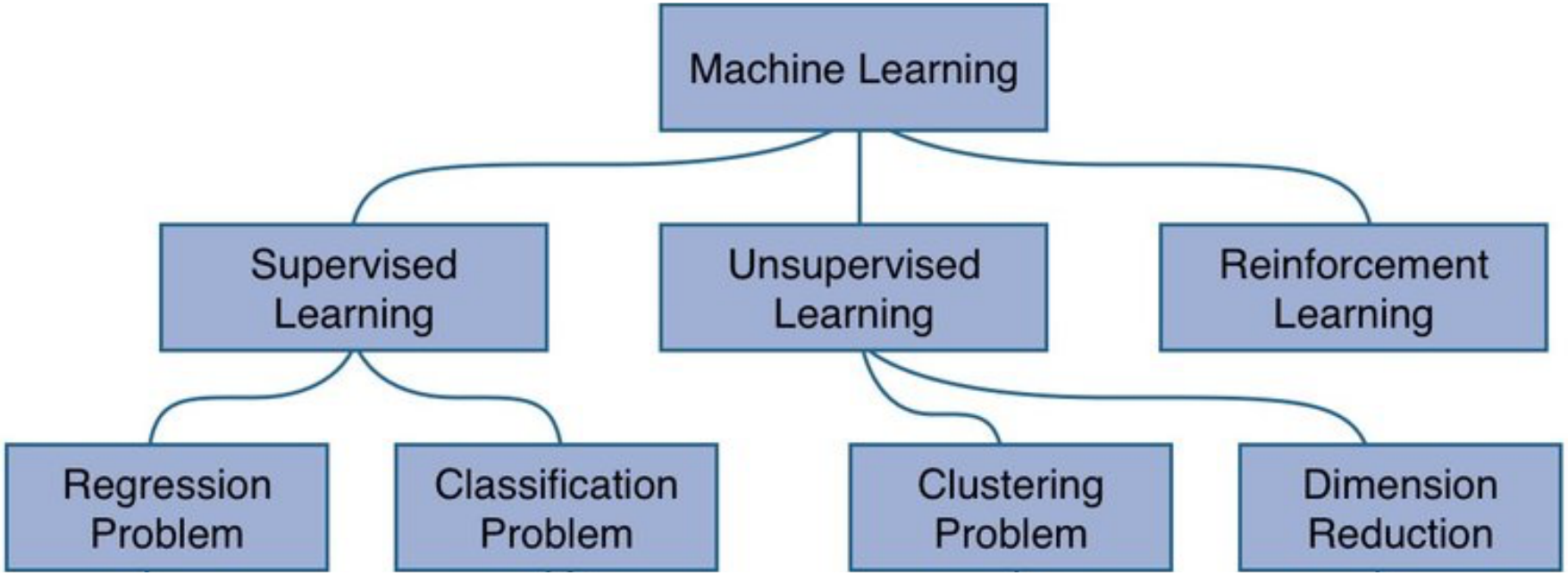
- 강의자료
  - https://github.com/sejongresearch/KISTI-Sejong-WinterSchool

기계학습 이론

기계학습 이론 간략 리뷰

강의 범주

- 본 강의에서 논의할 기계학습 범주



강의 범주

- 지도 학습의 대표적인 머신 러닝 방법
  - 분류 (classification)
  - 회귀 (regression)

- 분류
  - 분류는 미리 정의된 여러 클래스 레이블 중 하나를 예측하는 것
  - 두 개로만 나누는 이진 분류와 셋 이상의 클래스로 분류하는 다중 분류로 나뉨
  - 분류 예시: 얼굴 인식, 숫자 판별 등

- 회귀
  - 연속적인 숫자 또는 부동소수 점수(실수)를 예측하는 것
  - 회귀 예시: 주식 가격을 예측하여 수익을 내는 알고리즘 등

강의 범주

- 지도 학습의 대표적인 머신 러닝 방법
  - 분류 (classification)
  - 회귀 (regression)

- 각 문제는 분류와 회귀 중 어디에 해당되는가?
  - 미세먼지 예보 등급 예측 문제
  - 김장철 배추 가격 예측 문제
  - 코로나 감염 환자 수 예측 문제
  - 검사 수치 기반으로 암의 양성/음성 예측 문제
  - 고객 정보 기반으로 은행 대출 가능 여부 예측 문제
  - 리뷰 문장 기반으로 긍정/부정 리뷰 예측 문제
  - 내년 서울시 인구 수 예측 문제

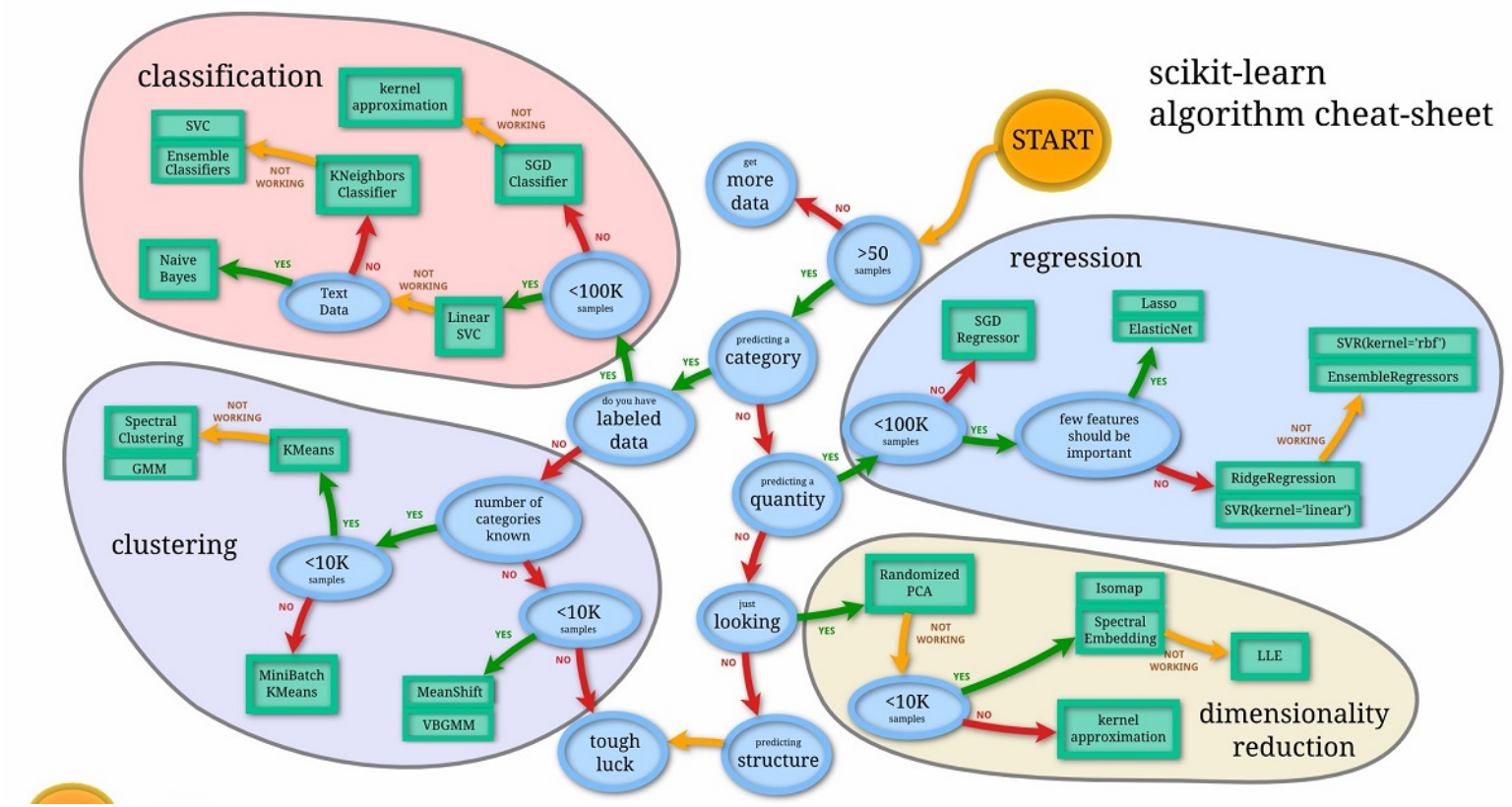
강의 범주

- 지도 학습의 대표적인 머신 러닝 방법
  - 분류 (classification)
  - 회귀 (regression)

- 각 문제는 분류와 회귀 중 어디에 해당되는가?
  - 미세먼지 예보 등급 예측 문제 ➔ 분류
  - 김장철 배추 가격 예측 문제 ➔ 회귀
  - 코로나 감염 환자 수 예측 문제 ➔ 회귀
  - 검사 수치 기반으로 암의 양성/음성 예측 문제 ➔ 분류
  - 고객 정보 기반으로 은행 대출 가능 여부 예측 문제 ➔ 분류
  - 리뷰 문장 기반으로 긍정/부정 리뷰 예측 문제 ➔ 분류
  - 내년 서울시 인구 수 예측 문제 ➔ 회귀

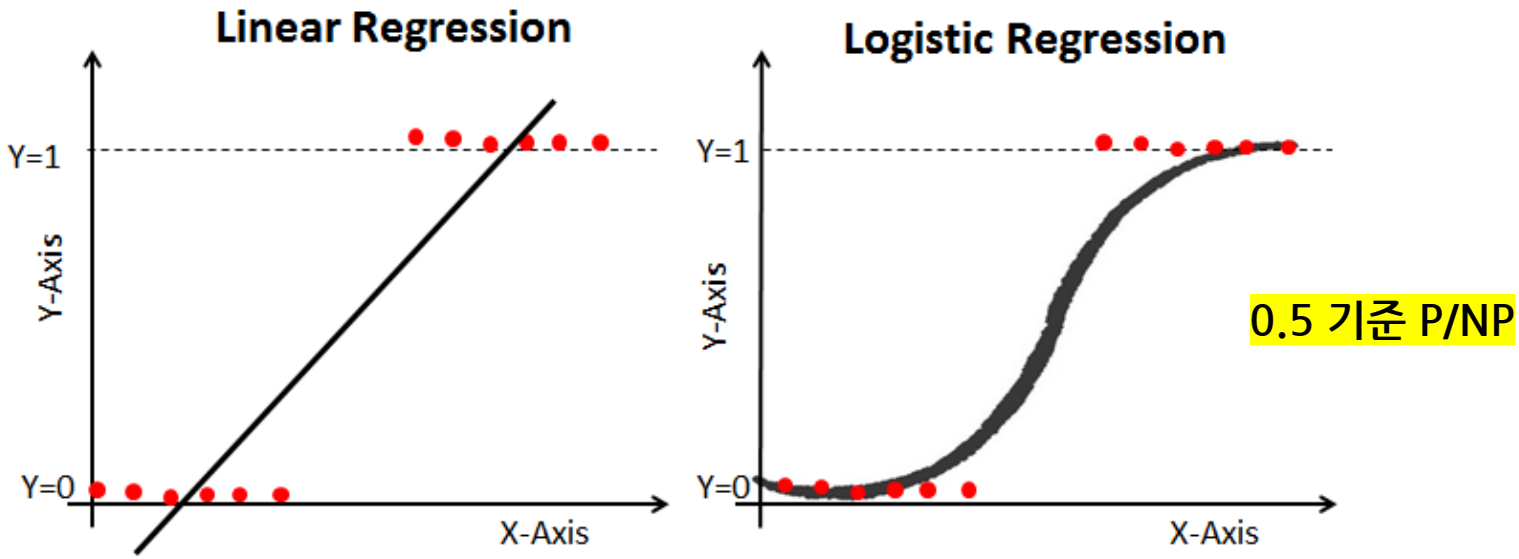
강의 범주

- 본 강의에서 실습할 기계학습 범주
  - Scikit-learn 을 이용한 실습
    - Scikit-learn : 파이썬으로 작성된 데이터 분석을 위한 범용 오픈 소스 라이브러리



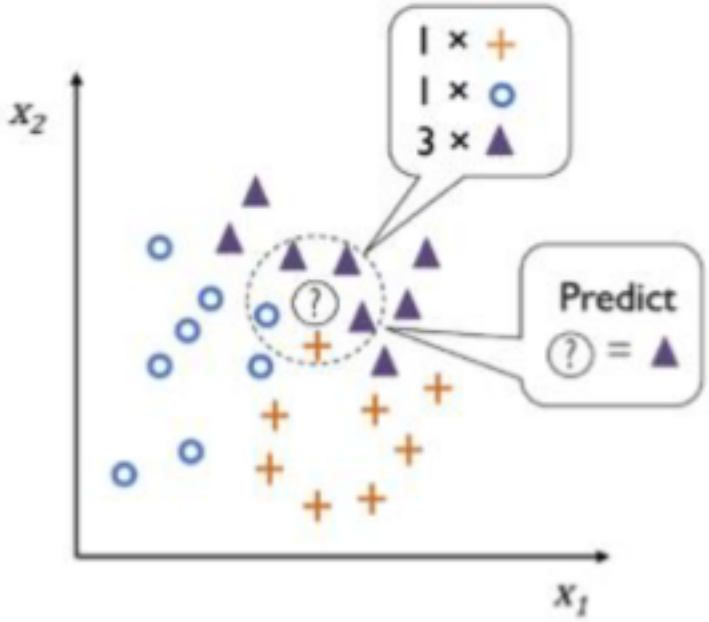
분류: LR (Logistic Regression)

- 로지스틱 회귀(Logistic Regression) 란?
  - 선형 회귀 분석과 다르게 종속 변수가 범주형 데이터를 대상으로 하며, 입력 데이터가 주어졌을 때 데이터의 결과가 특정 분류로 나뉘는 분류 기법에 해당함
  - 예시: 시험 공부 양(X축)에 따른 시험 통과/미통과(Y축) 여부 예측 문제



분류: KNN (K-Nearest Neighbor)

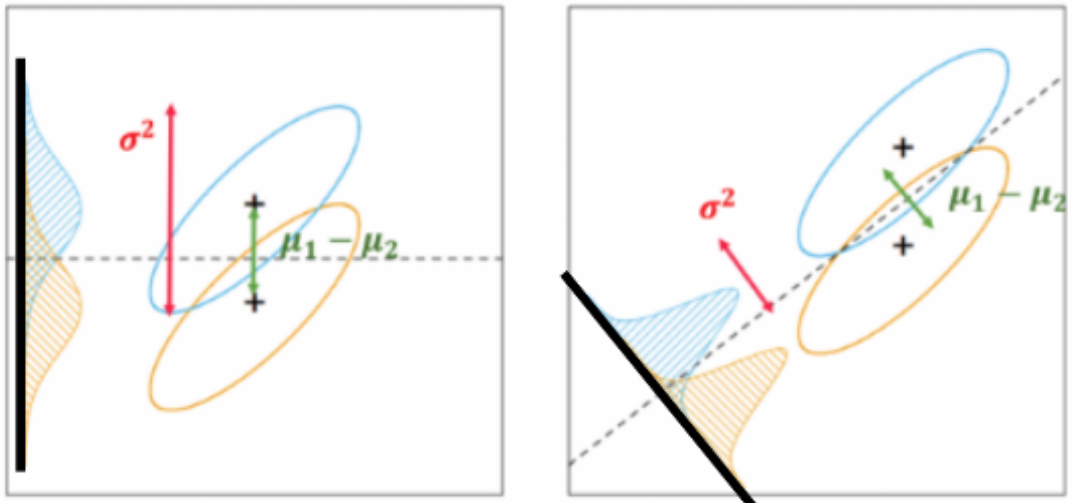
- KNN이란?
  - 주변 K개 데이터 클래스 중 가장 많은 클래스로 데이터를 분류하는 방법
  - 학습 데이터 자체가 모형일 뿐 모형의 파라미터를 추정하지 않는 방법
  - 매우 간단한 방법이지만 성능은 상대적으로 떨어지지 않는 방법



K=5인 KNN의 예시

분류: LDA(Linear Discriminant Analysis)

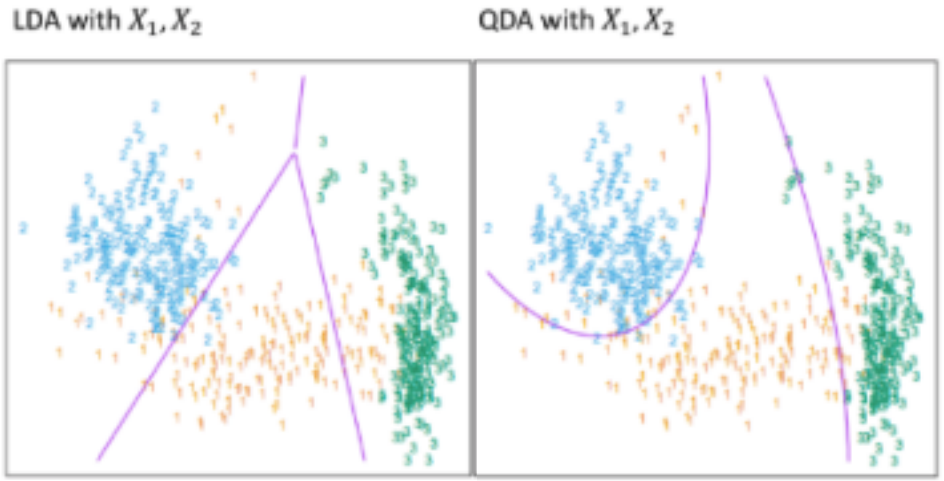
- 선형판별분석(LDA)란?
  - 데이터를 특정 한 축에 사영(projection)한 후에 두 범주를 잘 구분할 수 있는 직선을 찾는 방법
  - 결정 경계는 각 클래스 집단의 평균 차이는 크고, 분산은 작은 지점을 지정함
  - 각 클래스 집단은 비슷한 형태의 공분산 구조를 가진다고 가정함



LDA 결정 경계 후보 #1 LDA 결정 경계 후보 #2

분류: QDA(Quadratic Discriminant Analysis)

- 이차판별분석(QDA)란?
  - 데이터를 특정 한 축에 사영(projection)한 후에 두 범주를 잘 구분할 수 있는 곡선을 찾는 방법
  - 결정 경계는 각 클래스 집단의 평균 차이는 크고, 분산은 작은 지점을 지정함
  - 각 클래스 집단의 공분산 구조가 상이할 경우 사용하는 방법

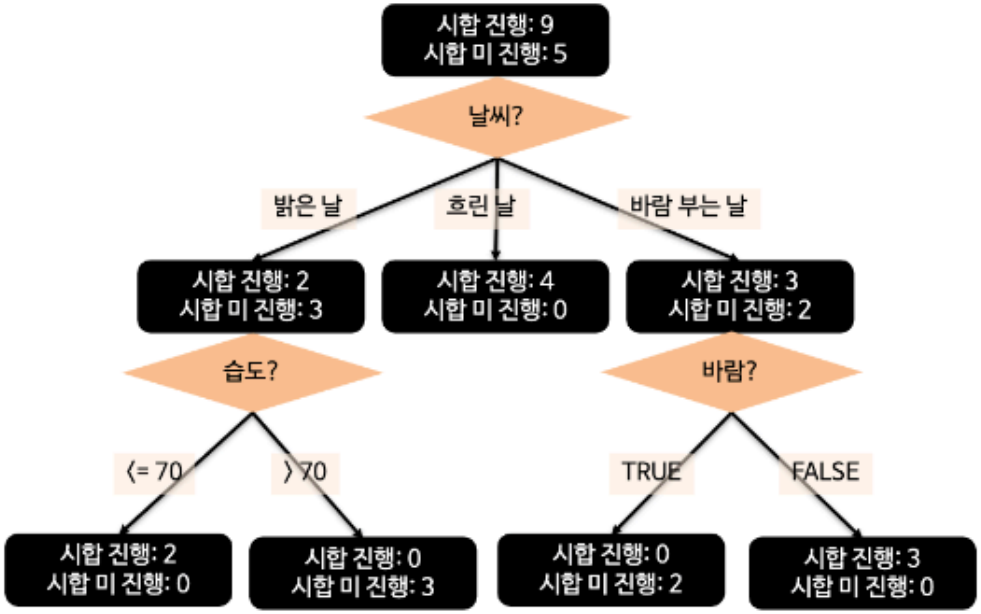


클래스 별 같은 공분산 구조를 가진다고 가정

클래스 별 다른 공분산 구조를 가진다고 가정

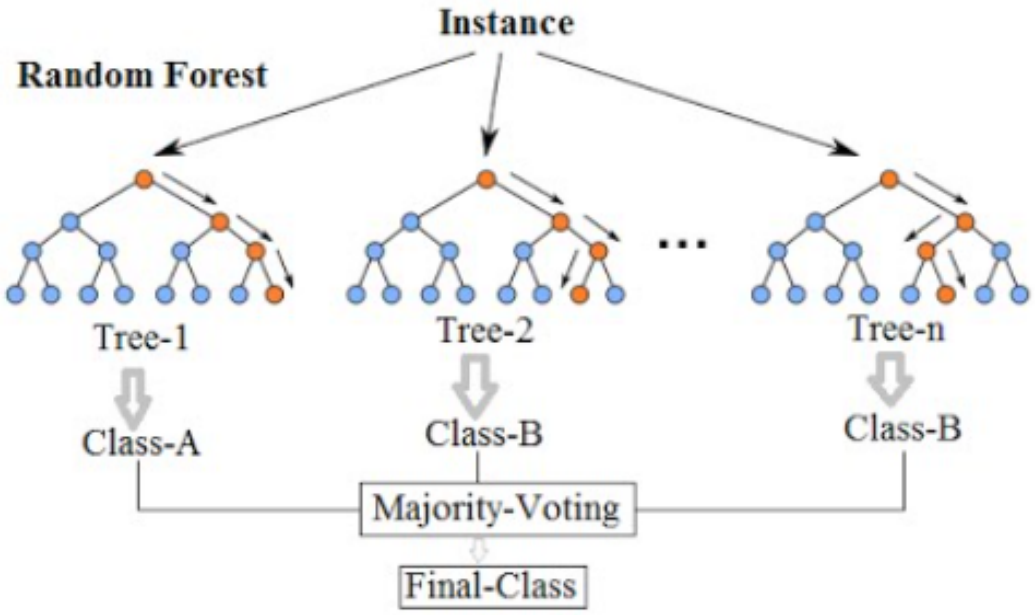
분류: DT (Decision Tree)

- 결정나무(Decision Tree) 란?
  - 의사 결정 규칙과 그 결과들을 트리 구조로 도식화한 의사결정 지원 도구
  - 개념적으로 질문을 던져 대상을 좁혀 나가는 "스무고개" 놀이와 비슷한 개념
    - 목적(Y)과 자료(X)에 따라 적절한 분리 기준과 정지 규칙을 지정하여 의사 결정 나무를 생성
  - 예시) 오늘의 야구 경기 진행 여부 예측 문제



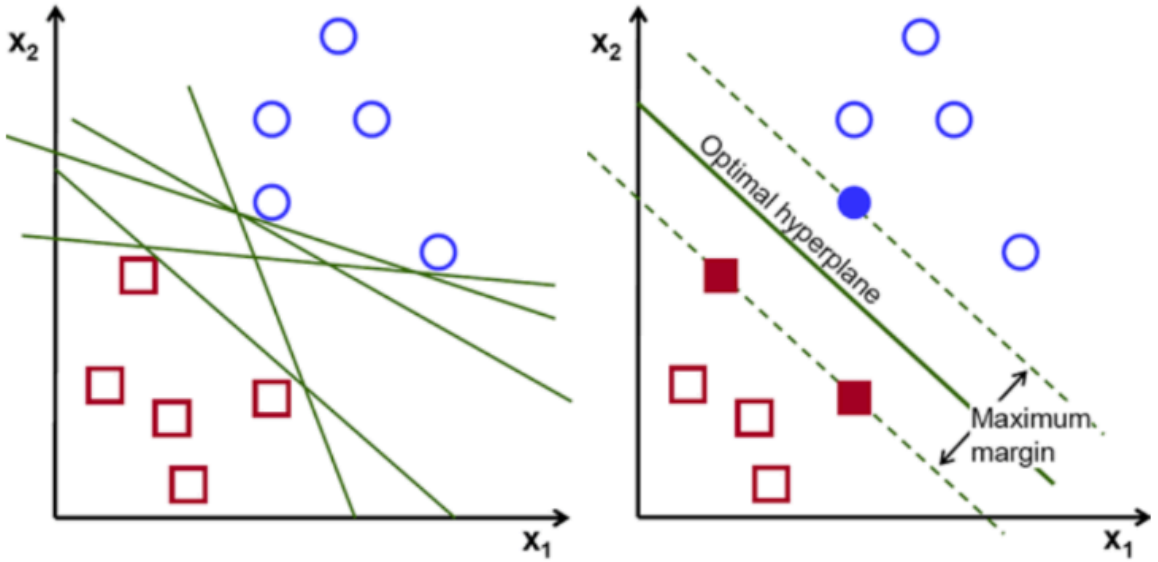
분류: RF (Random Forest)

- 랜덤 포레스트 (Random Forest) 란?
  - 분류 및 회귀 등에 사용되는 앙상블 학습 방법의 일종
    - 앙상블 학습법이란? 학습 알고리즘들을 따로 쓰는 경우에 비해 더 좋은 예측 성능을 얻기 위해 다수의 학습 알고리즘을 사용하는 방법
  - 훈련과정에서 구성한 다수의 결정 트리로부터 분류 또는 회귀 분석을 출력함



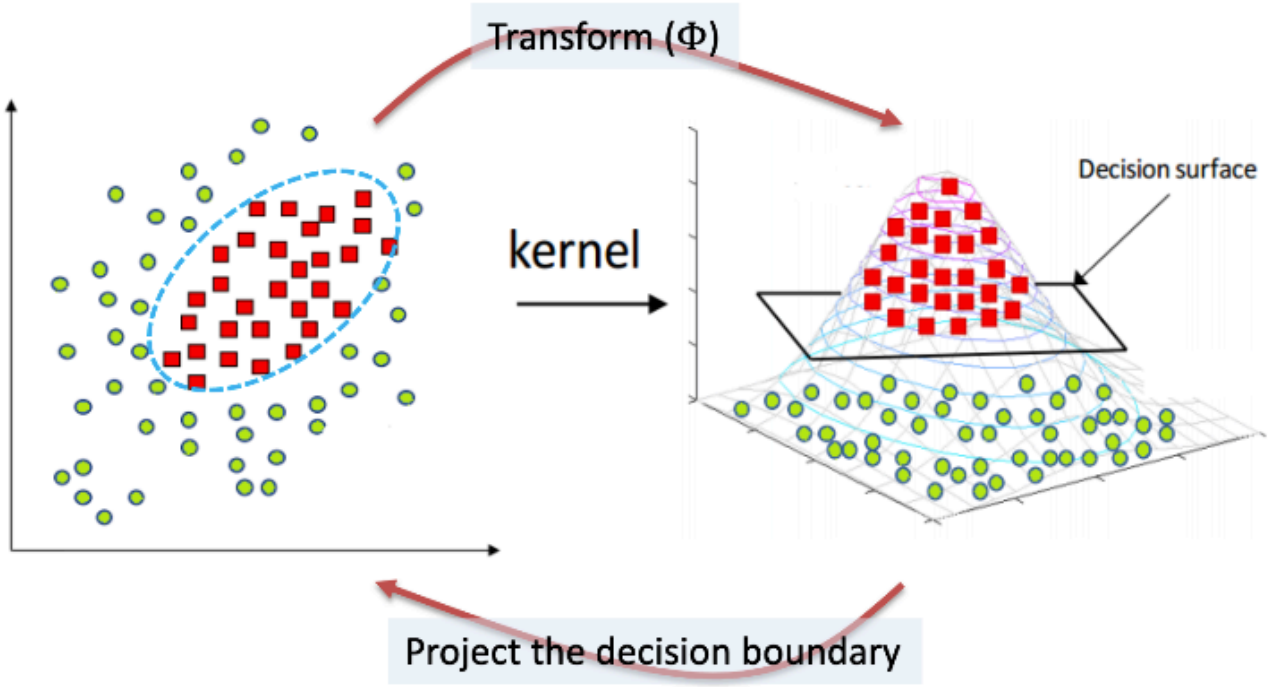
분류: SVM (Support Vector Machine)

- 선형 SVM이란?
  - 널리 사용되는 기계학습 방법론
  - 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, 주어진 데이터 집합을 바탕으로 새로운 데이터가 어느 카테고리에 속할지 판단하는 방법론
  - 결정 경계는 주변 데이터와의 거리(Margin)가 최대가 되도록 학습됨



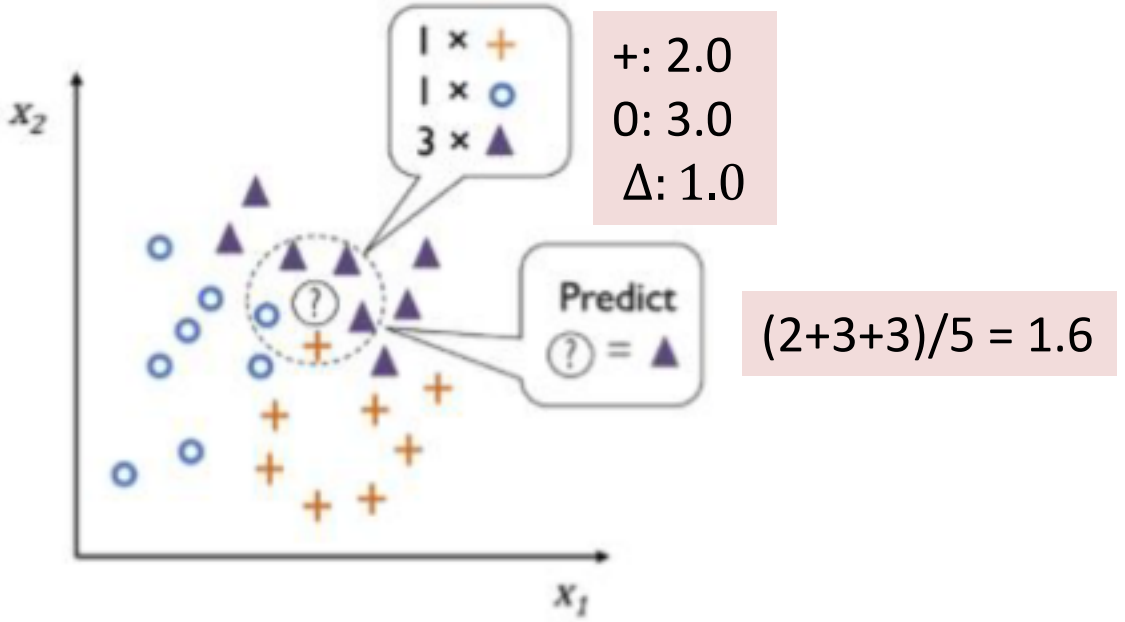
분류: SVM (Support Vector Machine)

- 비선형 SVM이란?
  - 커널(Kernel) 트릭을 적용한 SVM
  - 데이터를 선형으로 분류하기 위해 차원을 높이는 방법을 사용
  - Feature Map(Φ)을 통해 차원을 높임. 즉, X대신 Φ(X)를 사용



회귀: KNN

- **KNN 회귀란?**
  - 회귀 역시 분류와 원리는 동일
  - 주변 K개 이웃 데이터의 평균으로 결정
  - 학습 데이터 자체가 모형일 뿐 모형의 파라미터를 추정하지 않는 방법



회귀: DT (Decision Tree)

- 회귀 나무(Regression Tree) 란?
  - 목표 변수가 연속형 변수일 경우, 의사결정나무는 회귀나무라고 함
  - 회귀 알고리즘은 실제 값과 예측 값의 평균 차이가 작도록 트리를 생성함
  - 예측 데이터의 회귀 값은 끝 마디 집단의 평균값으로 결정함