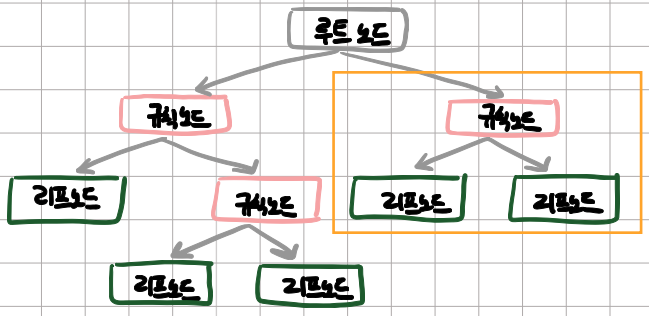
Week 2 - Ch4. 분류

01 분류의 개요

- 분류: 학습 데이터로 주어진 데이터의 피처와 레이블값을 머신러닝 알고리즘으로 학습해 모델 생성
  → 모델에 새로운 데이터 값이 주어졌을 때 미지의 레이블값 예측

02 결정 트리

- 결정트리 (Decision Tree)



- 규칙노드 : 규칙조건이 됨
- 리프노드 : 결정된 클래스 값
- 서브트리 : 새로운 규칙 조건마다 생성

- 트리를 어떻게 분할 할것인가?
  ⇒ "균일도" 최대는 균일한 데이터 세트를 구성할수 있도록 분할
  - 인트로피 → 정보이득지수
  - 지니계수

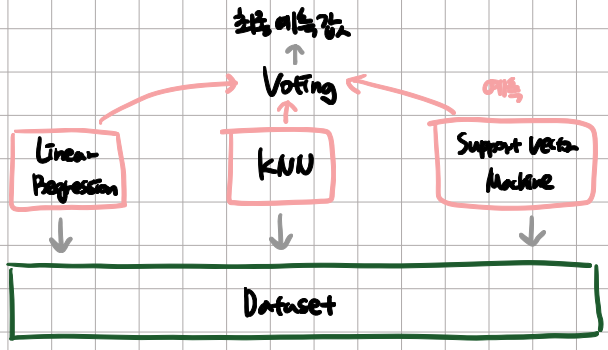
- 결정트리의 특징
  - 장점 — 쉽다, 직관적이다
    — 사전가공 영향도 ↓
  - 단점 : overfitting ~ 알고리즘 성능↓ → 트리크기 제한~튜닝

- 결정 트리 파라미터
  CART - Decision Tree Classifier 클래스
  - min_samples_split
  - min_samples_leaf
  - max_features
  - max_depth
  - max_leaf_nodes

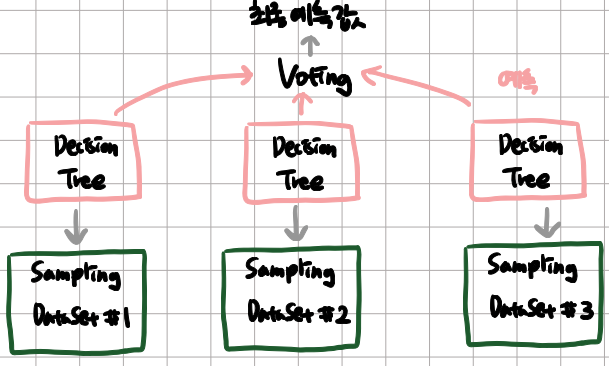
03 앙상블 학습

- 앙상블 학습 (Ensemble Learning) :
  여러개의 분류기 (Classifier) 를 생성, 그 예측 결합
  → 보다 정확한 최종 예측 도출하는 기법
  — 정형 데이터 분류 ~ 뛰어난 성능.

- 보팅 (Voting) :
  서로 다른 알고리즘을 가진 분류기 결합



- 배깅 (Bagging) :
  같은 알고리즘 기반의 분류기, 데이터 샘플링을 다르게



↳ 부트스트래핑 (Bootstrapping) 분할 방식.
CV와 다르게 dataset 간 중첩허용O

- 부스팅 (Boosting) :
  앞에서 학습한 분류기 ~ 예측 틀림 → 다음 분류기에
  가중치 부여하면서 학습, 예측 진행

- 하드 보팅 (Hard Voting) : 다수결로 결정
  소프트 보팅 (Soft Voting) : 분류기의 레이블값 결정 확률을 평균내서 가장 높은 값으로

04 랜덤 포레스트

- 랜덤 포레스트 : 배깅의 대표적인 알고리즘. 결정트리를 기반으로.
- RandomForestClassifier
  - n_estimators
  - max_features
  - max_depth, min_samples_leaf 등 결정트리에 사용되는 파라미터