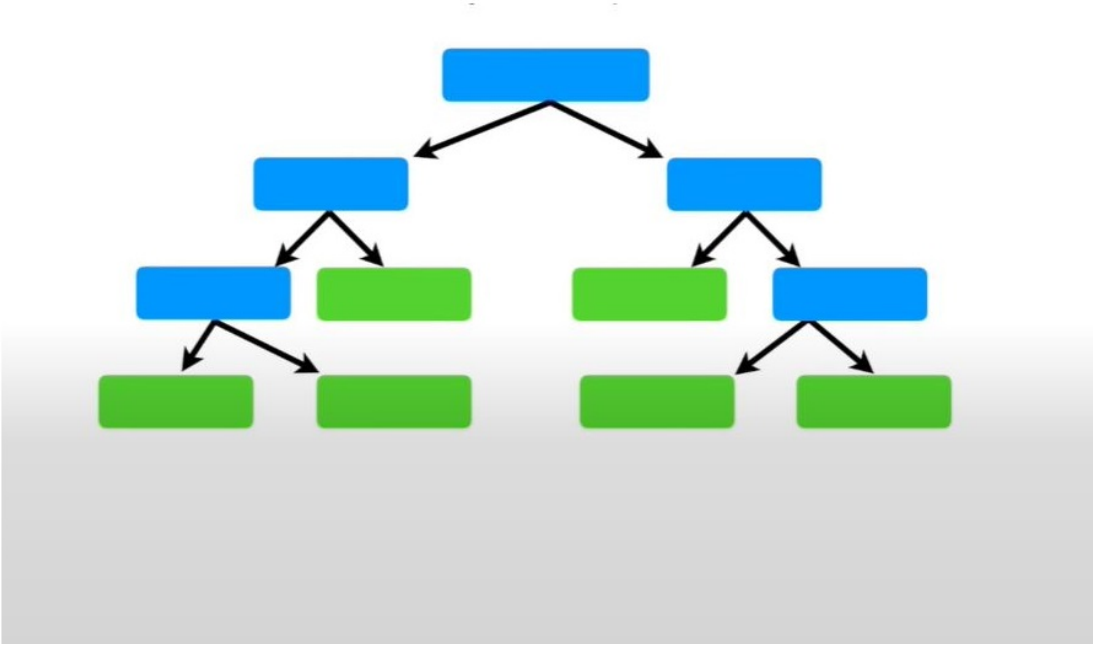
집단지성을 활용한 분류 알고리즘

Classification with Random Forests

2020학년도 2학기

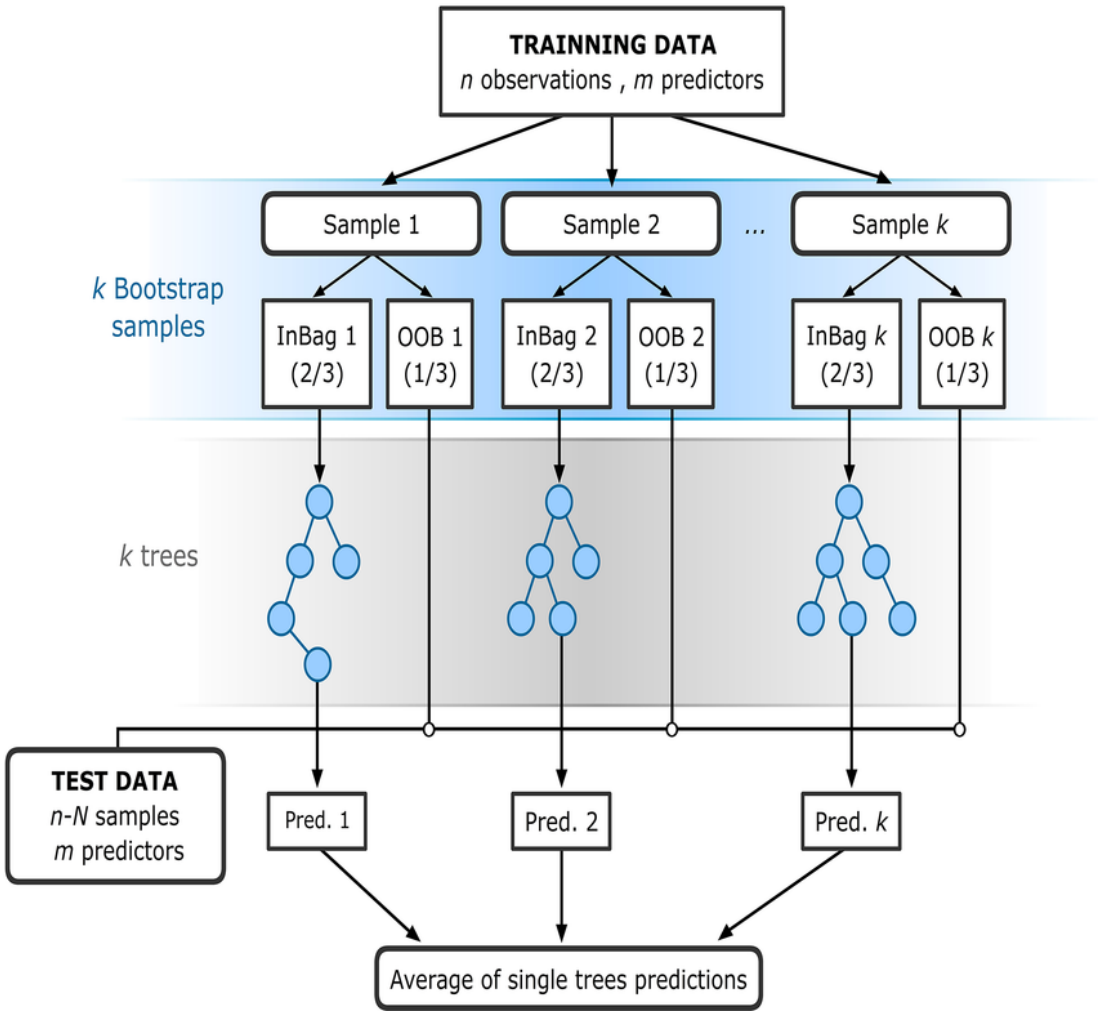
Decision Trees 복습



Decision Trees

- 아이디어
  - 불순도가 낮아지도록 하는 분류 기준(→node)을 순차적으로 적용하기
- 단점
  - 새로운 데이터에 대한 예측력이 뛰어나지는 않은 것으로 알려져있다.
  - "Trees have one aspect that prevents them from being the ideal tool for predictive learning, namely inaccuracy. They seldom provide predictive accuracy comparable to the best that can be achieved with the data at hand" (Hastie, 2009)

Random Forests



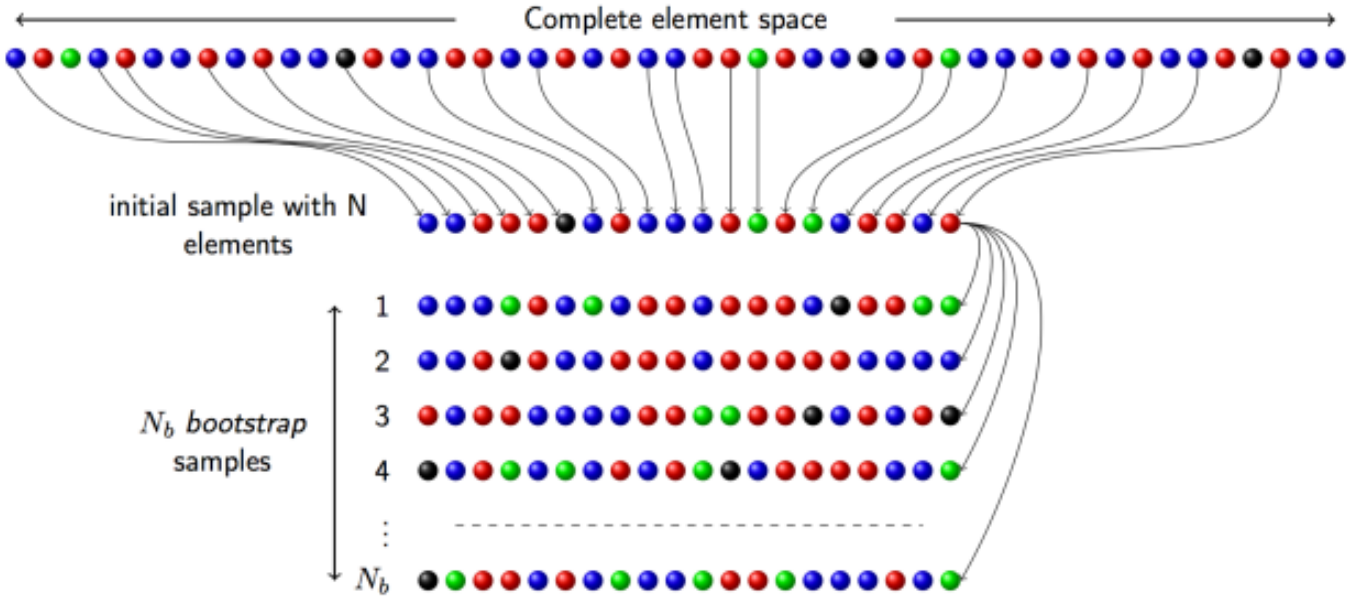
Random Forests 개요

- 아이디어
  - 데이터를 재구성한 후, q 개의 임의(random)의 변수만을 node의 후보로 선택하여 Decision Tree를 만든 후, 이러한 과정으로 수많은 나무(forests)를 만들어서 다수결에 의한 분류 결과를 산출하는 알고리즘
- 순서
  - 1. 데이터 재구성하기
  - 2. q 개의 특성만을 후보로 삼아 나무 만들기
  - 3. 앞의 방식을 다시 적용하여 수많은 나무를 만들기
  - 4. 수많은 나무들의 다수결에 의해 분류하기
  - 5. 데이터 재구성 시 누락된 데이터 하나하나로 부분적 예측력 집계하기
  - 6. 최적의 q값을 정하여 예측력을 최대화하기

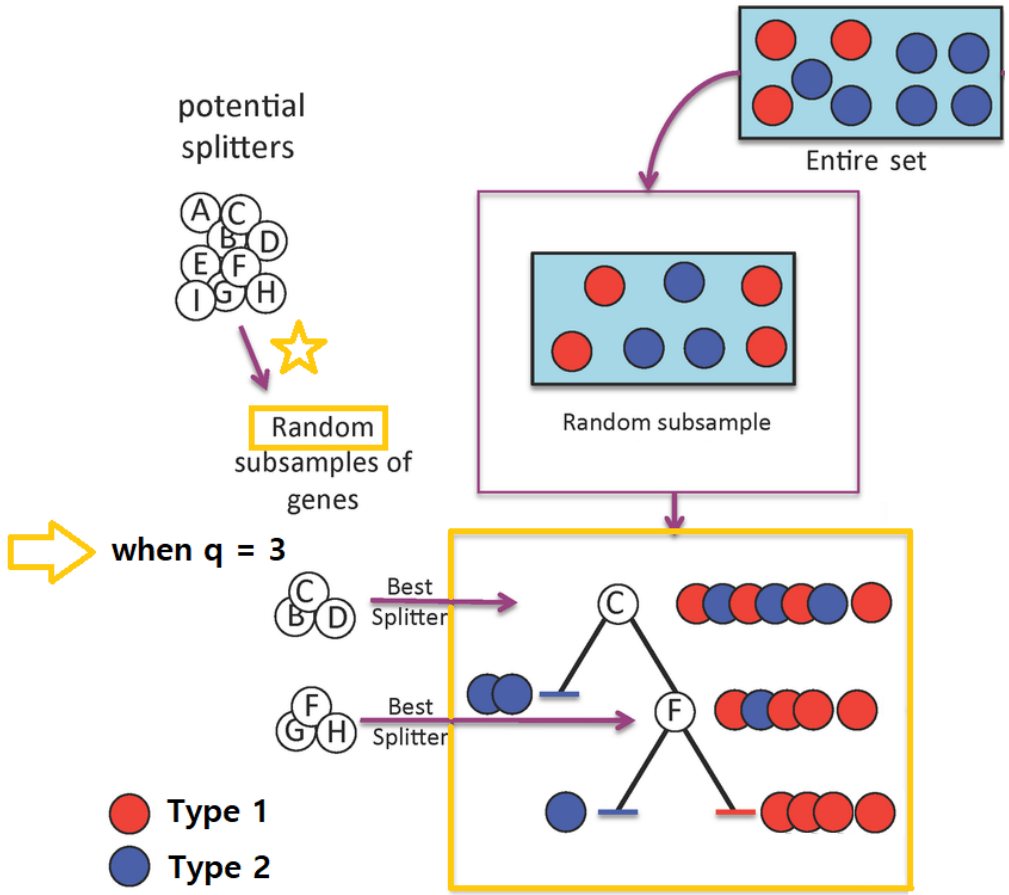
데이터 재구성하기(Bootstrapping)

- Bootstrapping
  - Random sampling with replacement
  - 기존의 샘플을 활용하여 가상의 샘플을 구성한다.
    (표본의 크기는 기존의 것보다 작거나 같도록 하는 것이 일반적이다.)
  - 가상의 샘플을 구성할 때, 중복추출을 허용한다.
  - 나무가 총 k그루인 숲을 만들고 싶다면 k개의 가상 샘플을 만든다.
  - 가상의 샘플 안에는 기존의 샘플 중에서 누락되는 데이터도 있기 마련이다.

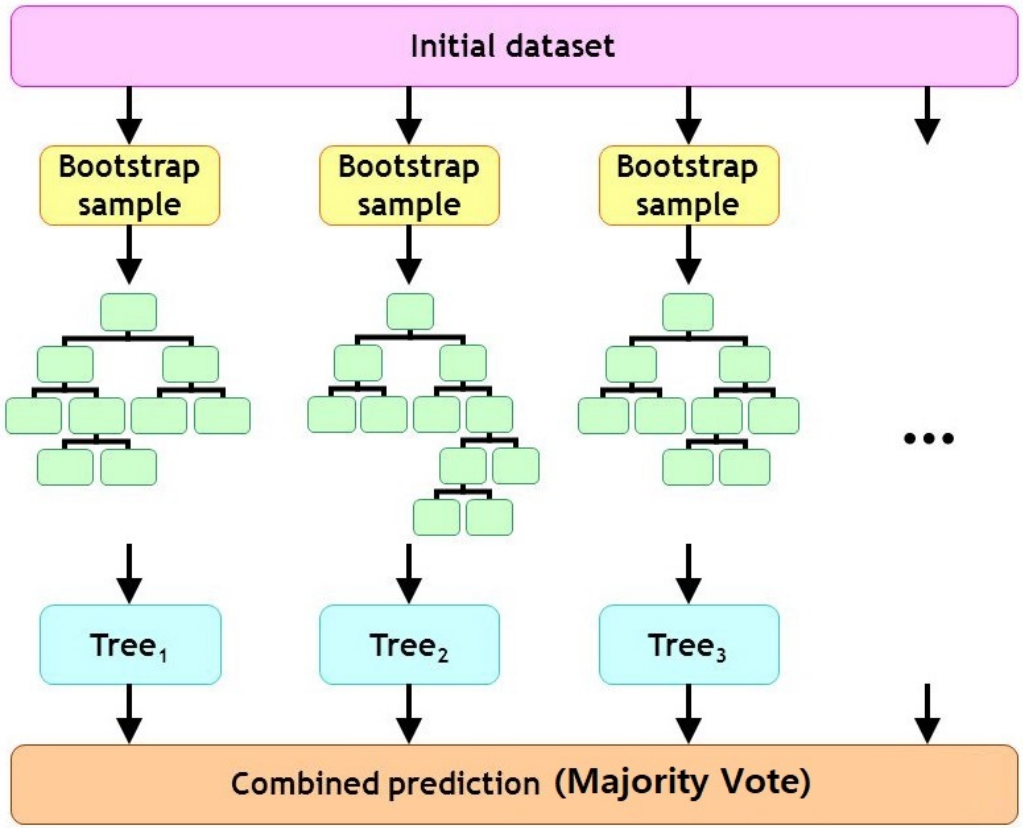
Bootstrapping ($k = N_b$)



q 개의 특성만을 후보로 삼아 나무 만들기



수많은 나무를 만들어 다수결로 분류하기



누락된 데이터로 모델의 성능을 개선하기

- Out-of-the-Bag Sample
  - Bootstrapping을 할 때 누락된 데이터들을 Out-of-the-Bag Sample(OOB 샘플)이라고 부른다.
  - k그루의 나무를 만드는 과정에서 한 번이라도 누락된 적이 있는 OOB 샘플이 총 t개 있다고 하자.

누락된 데이터로 모델의 성능을 개선하기

- Out-of-the-Bag Sample
  - i번째 OOB 샘플은 총 k그루의 나무 중에서 s그루의 나무를 만들 때 누락되었다고 한다면,
  - 해당 s그루의 나무에 대해서 i번째 OOB샘플로 예측력 테스트를 할 수 있다.
  - 그렇게 하여 측정된 총 s개의 예측력을 평균한 것이 i번째 OOB 예측력이다.

누락된 데이터로 모델의 성능을 개선하기

- Out-of-the-Bag Sample
  - 이와 같은 방식으로 총 t번 OOB 예측력을 계산할 수 있고,
  - t개의 OOB 예측력에 평균을 취한 것이 현재(q=3) 랜덤 포레스트 모델의 성능이라 할 수 있다.
  - q값을 바꾸어가면서 위와 같은 과정을 반복하면, 최적의 q값을 찾을 수 있다.