필기 과제

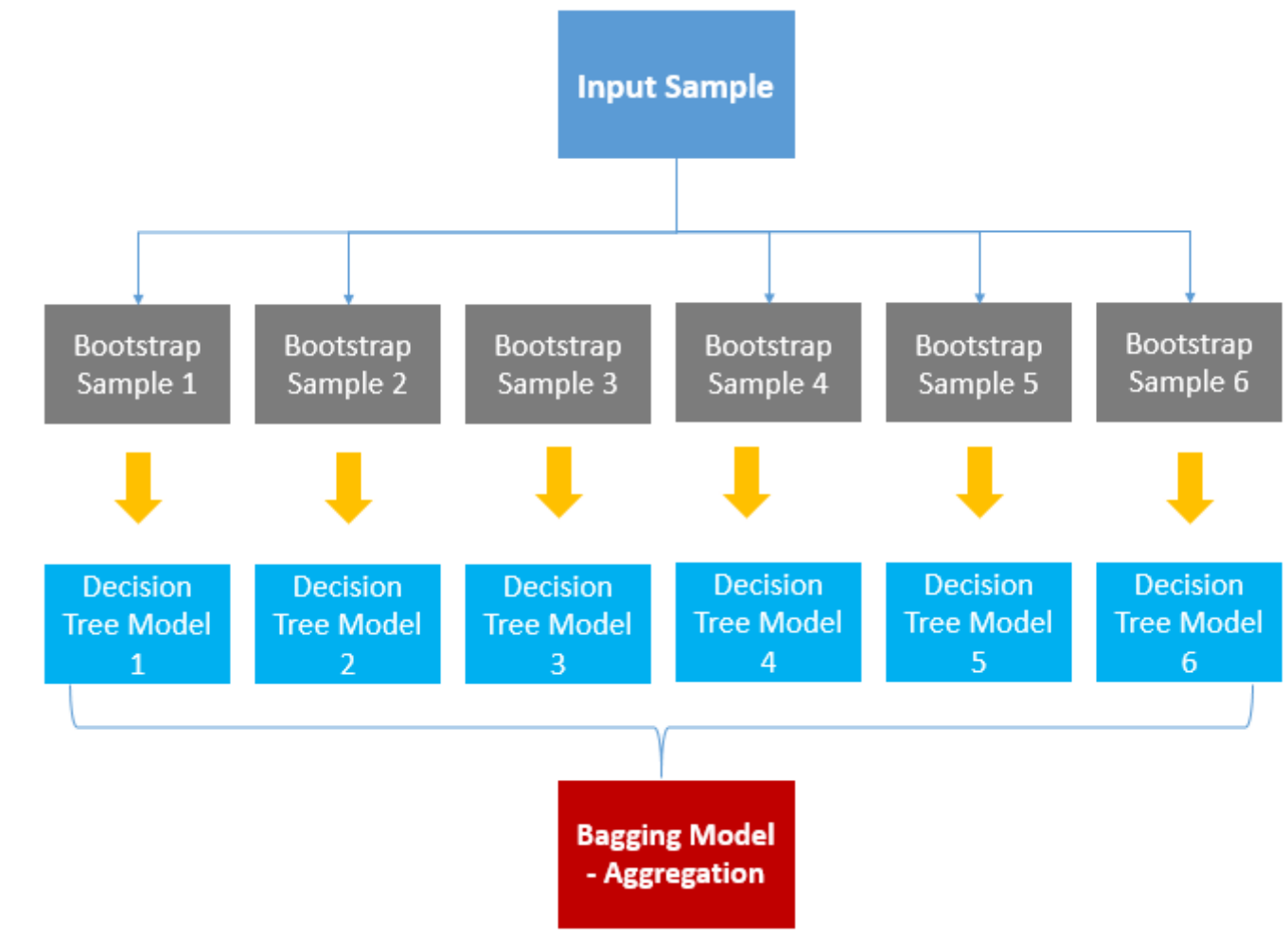
분산과 편차에 따른 모델 복잡도

- **Bias** : 학습된 분류기와 실제 값 사이의 제곱 에러, 정확도와 비슷한 개념
- **Variance** : 학습된 분류기들이 각기 다른 학습셋에 성능의 변화정도가 급하게 변하는지 안정적으로 변하는지를 나타내는 척도

  → 예측값들이 정답과 대체로 멀리 떨어져 있으면 편향이 높다
  예측값들이 자기들끼리 뭉쳐있지 않으면 분산이 높다
  - 일반적으로 한쪽이 증가하면 한쪽이 감소하는 경향
- Bias가 높고 Var가 낮음 → 모델 복잡도 높음
  : 모델이 매우 안정적인 결과를 내놓으며 데이터가 달라도 비슷한 구역으로 예측값을 매핑. 하지만 비슷한 구역이 잘못된 위치일 수 있음
- Bias가 낮고 Var가 높음 → 모델 복잡도가 낮음
  : 모델이 **높은 variance**와 **낮은 bias**를 가질 때, 일부는 정확하게 매핑되지만 많은 데이터가 정확하게 예측하지 못함
  데이터가 조금만 달라져도 완전히 다른 결과가 나올 수 있음

배깅과 부스팅모델 각각의 개념과 차이점

- Bagging (Bootstrap Aggregation)
  : 샘플을 여러 번 뽑아 각 모델을 학습시켜 결과물을 집계하는 방법
  ex) RandomForest

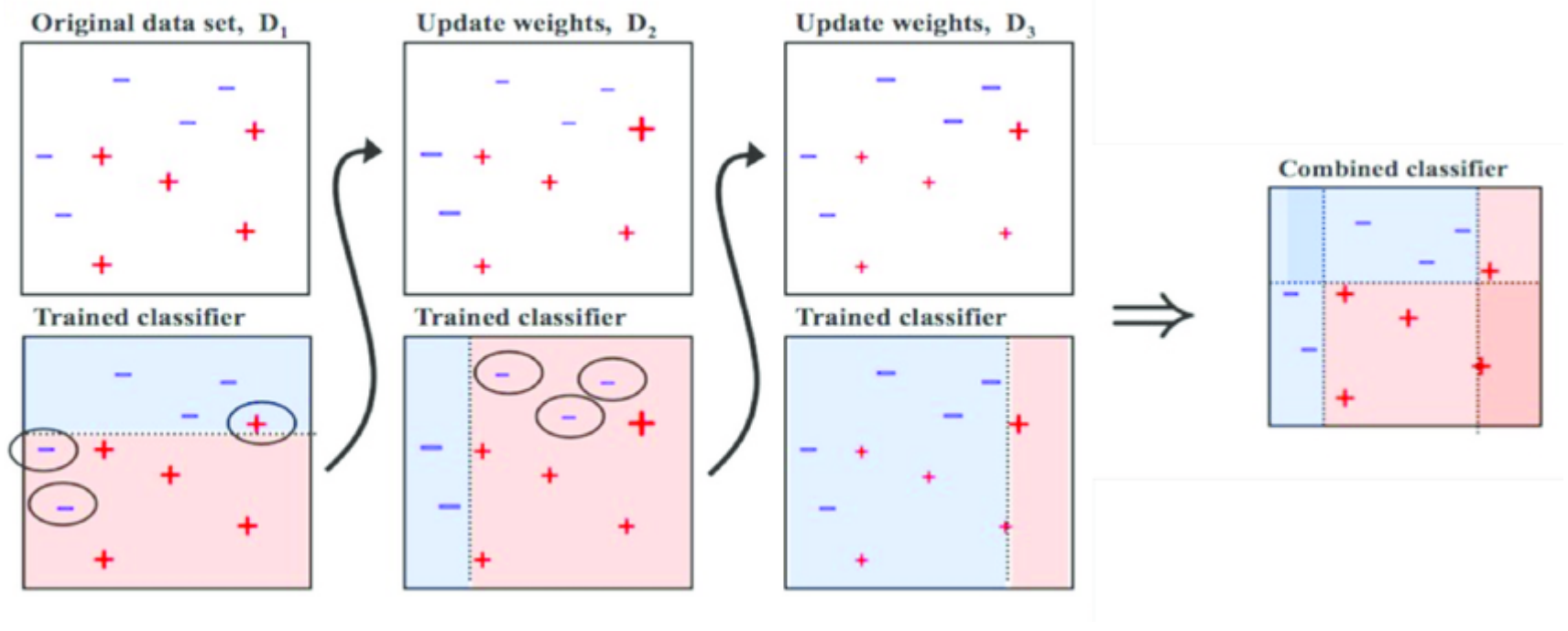


  - 데이터에서 샘플을 여러개 추출 (복원 랜덤 샘플링)하여 의사결정 나무를 생성
  - 예측 모델의 분산이 클 때 분산을 감소시키기 위해 사용
  - 분류 모델일 때, 각 트리의 결과값의 **다수결**로 최종 분류

    회귀 모델일 때, 각 트리의 결과값의 **평균값**으로 최종 분류

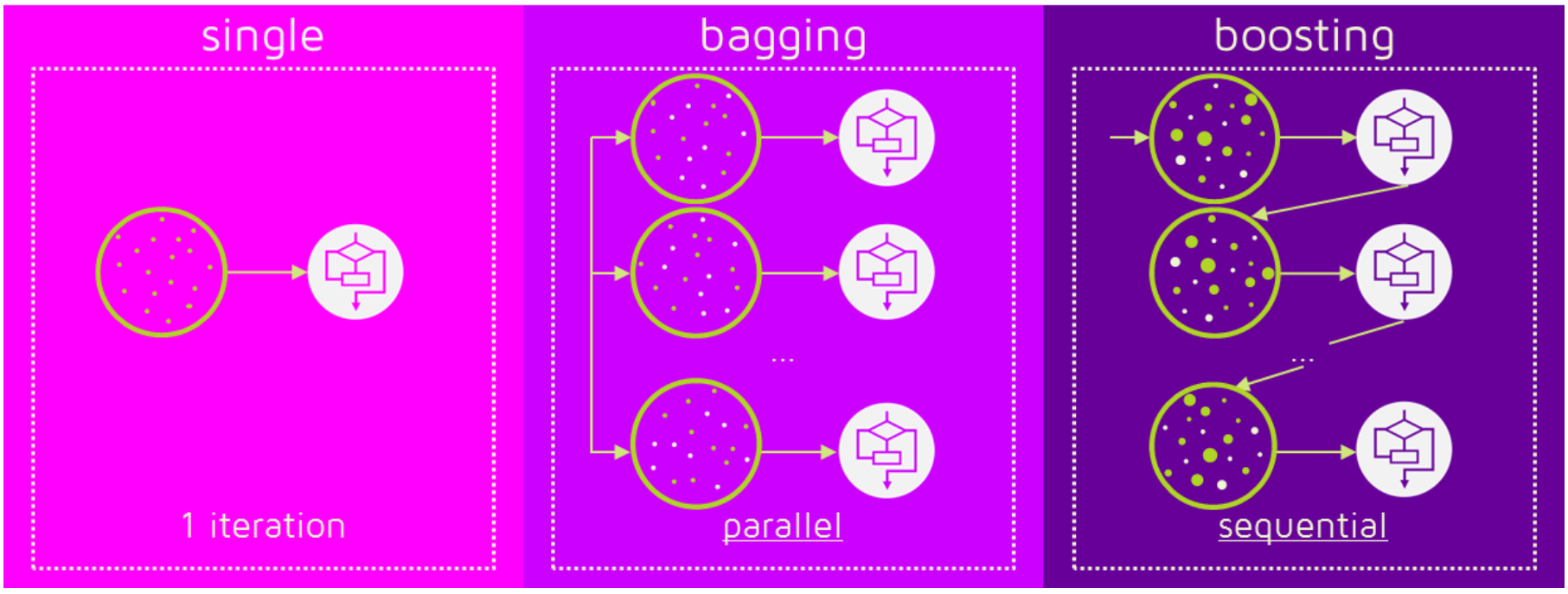
- Boosting

: 가중치를 활용하여 약 분류기를 강 분류기로 만드는 방법
ex) XGBoost, GBM, CatBoost



- 각 회차에 맞추지 못한 데이터에 가중치를 부여하여 다음 학습 모델에 반영 → 예측력 상승
- 예측 모델의 편차가 클 때 편차를 감소시키기 위해 사용

- 차이점



- 간단하게, 배깅은 각각의 의사결정 나무가 독립적인 존재이기에 여러 개의 독립적인 결정 트리가 각각 값을 예측한 뒤, 그 결과 값을 집계해 최종 결과 값을 예측하는 방식으로 진행됨.

  하지만, 부스팅의 경우에는 처음 모델이 예측을 하면 그 예측 결과에 따라 데이터에 가중치가 부여되고, 부여된 가중치가 다음 모델에 영향을 주기 때문에 이 나무들이 연결되어 있다고 볼 수 있음.