Part. 04

Ensemble Learning

RandomForest

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택

I Ensemble Learning의 종류

▪ Ensemble Learning의 종류

- Bagging : 모델을 다양하게 만들기 위해 데이터를 재구성
- RandomForest : 모델을 다양하게 만들기 위해 데이터 뿐만 아니라, 변수도 재구성
- Boosting : 맞추기 어려운 데이터에 대해 좀더 가중치를 두어 학습하는 개념
  Adaboost, Gradient boosting (Xgboost, LightGBM, Catboost)

Tree기반의 단일 모델 (패키지 함수)

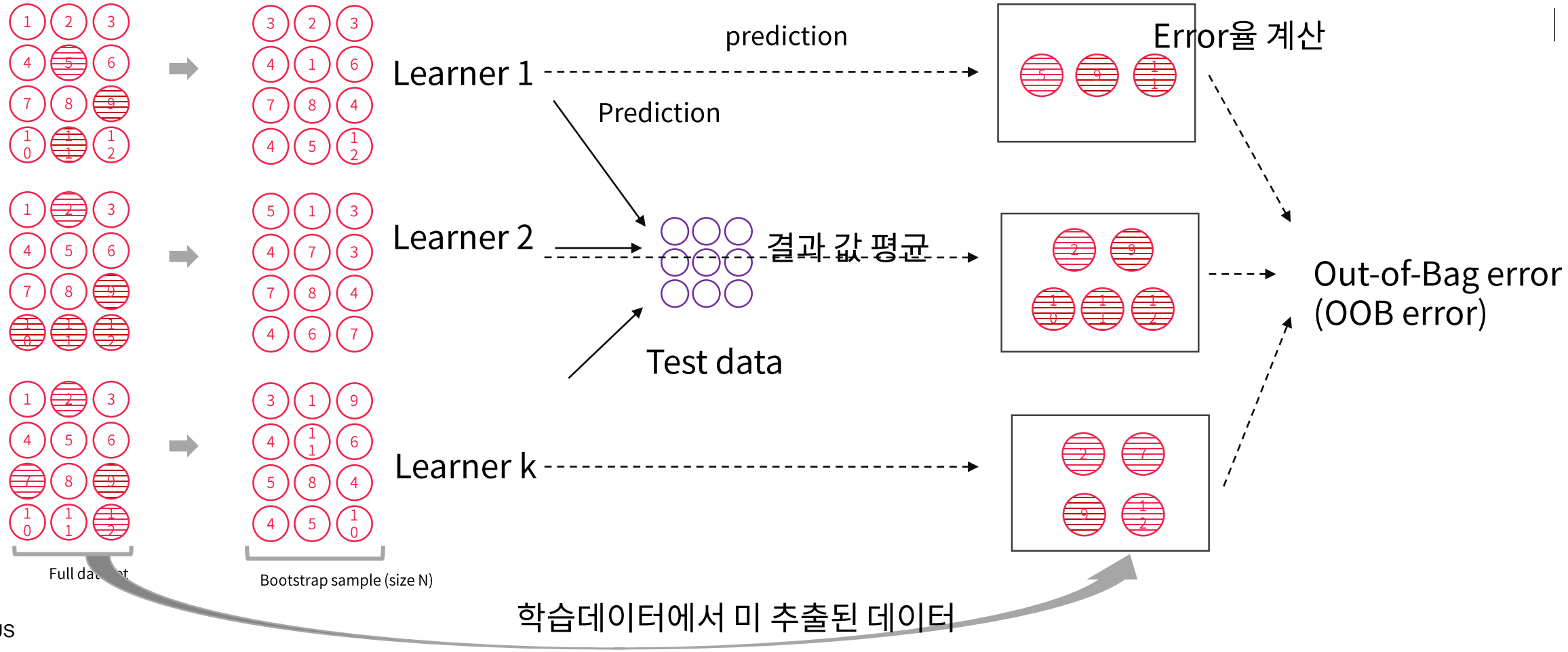
- Stacking : 모델의 output값을 새로운 독립변수로 사용

“Ensemble의 한 개념”

Copyright FASTCAMPUS Corp. All Rights Reserved

I Ensemble Learning의 종류

- **Bagging (bootstrap aggregating)**



Copyright FASTCAMPUS Corp. All Rights Reserved

I Ensemble Learning의 종류

- Bagging (bootstrap aggregating)

Tree vs Bagging

깊이 성장한 트리 : 분산 증가, 편향 감소

Bagging : 트리들의 편향 유지, 분산 감소 / 학습데이터의 noise에 강건해짐 / 모형해석의 어려움

Copyright FASTCAMPUS Corp. All Rights Reserved

I Ensemble Learning의 종류

- **Bagging (bootstrap aggregating)**

Bagging model(여러 트리들)의 분산은 각각 트리들의 분산과 그들의 공분산으로 이루어져있음

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

전체데이터에서 복원 추출하였으나, 각각의 트리들은 중복되는 데이터를 다수 가지고 있기 때문에 독립이라는 보장이없음

↓

$Cov(X,Y)$ = 0 이라는 조건을 만족하지 못함 (비슷한 tree가 만들어질 확률이 높음)

↓

Tree가 증가함에 따라 오히려 모델 전체의 분산이 증가 할 수도 있음

↓

각 Tree간 공분산을 줄일 수 있는 방법이 필요함

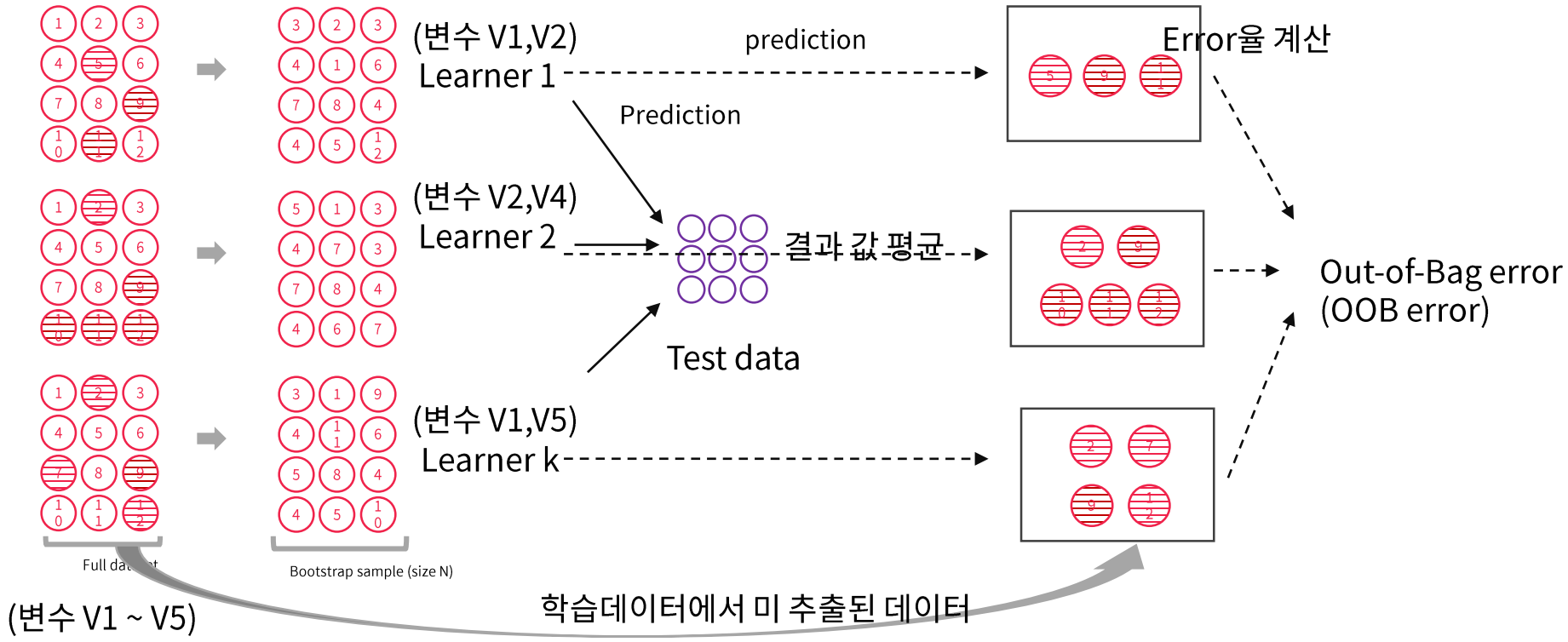
Copyright FASTCAMPUS Corp. All Rights Reserved

I Ensemble Learning의 종류

- **RandomForest**

  Ensemble learning의 기본 컨셉 - '다양한 모델 '

  데이터 뿐만이 아니라, 변수도 random하게 뽑아서 다양한 모델을 만들자(base learner간의 공분산을 줄이자)



Copyright FASTCAMPUS Corp. All Rights Reserved

I Ensemble Learning의 종류

- **RandomForest**

  Ensemble learning의 기본 컨셉 - '다양한 모델 '

  데이터 뿐만이 아니라, 변수도 random하게 뽑아서 다양한 모델을 만들자(base learner간의 공분산을 줄이자)

  뽑을 변수의 수는 hyper parameter (일반적으로 $\sqrt{p}$ 사용)

  모델의 분산을 줄여 일반적으로 Bagging보다 성능이 좋음

Copyright FASTCAMPUS Corp. All Rights Reserved

Part. 04

Ensemble Learning

Boosting

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택