Part. 04

Ensemble Learning

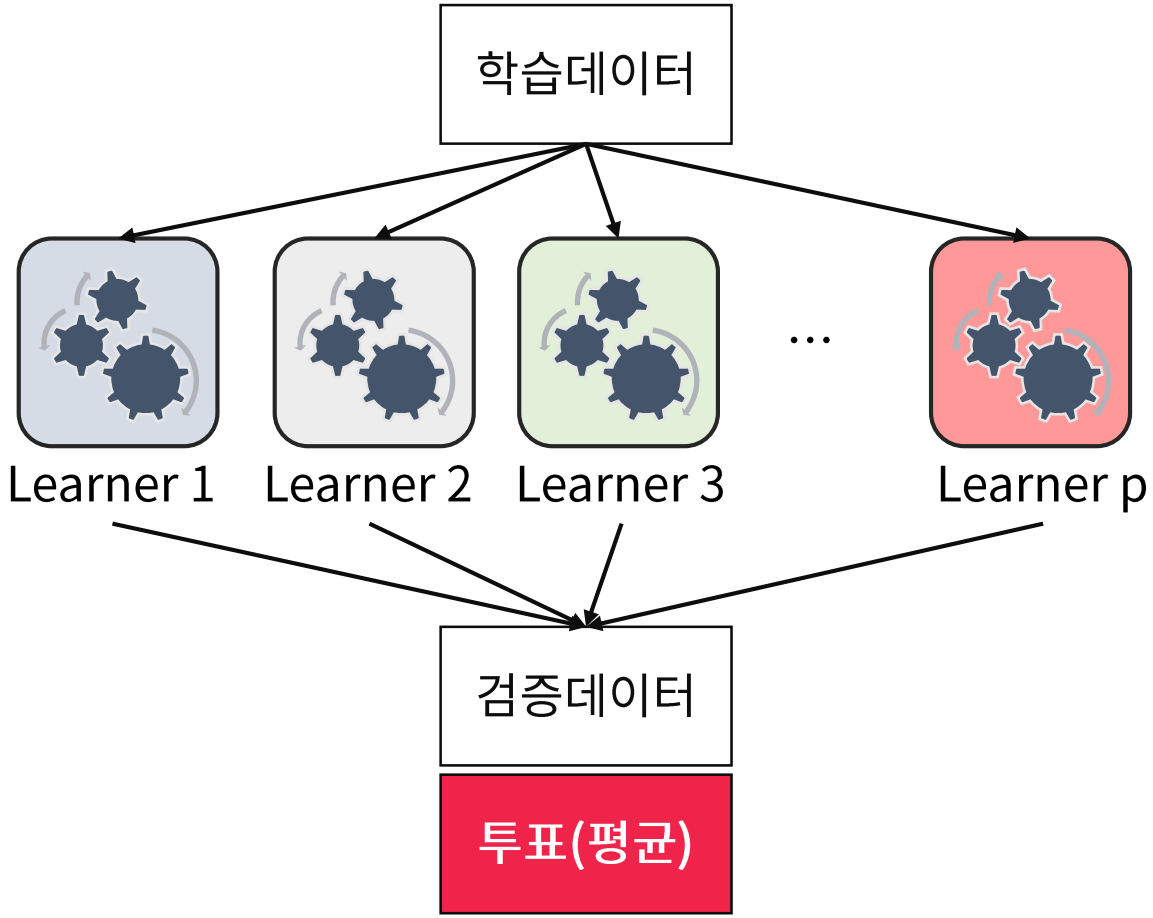
Ensemble의 기법 review

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택

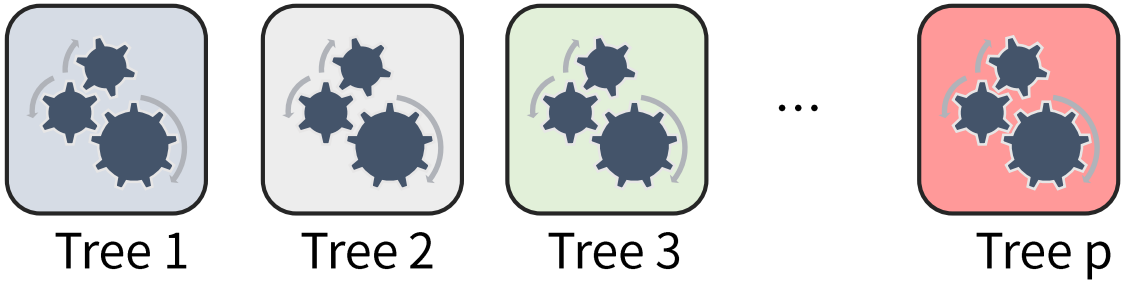
I Ensemble 기법 review

- **Ensemble Learning**
  - 여러 개의 기본 모델을 활용하여 하나의 새로운 모델을 만들어내는 개념



I Ensemble 기법 review

- **Ensemble Learning**
  - Test 데이터에대해 다양한 의견(예측값)을 수렴하기 위해 overfitting이 잘되는 모델을 기본적으로 사용



  - Ensemble개념 자체는 여러 모델의 조합을 뜻하기 때문에 Tree가 아닌 다른 모델을 사용해도 무방
  - 가장 많이 쓰이는 RandomForest, Boosting은 이 Tree기반 모델

I Ensemble 기법 review

▪ 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)

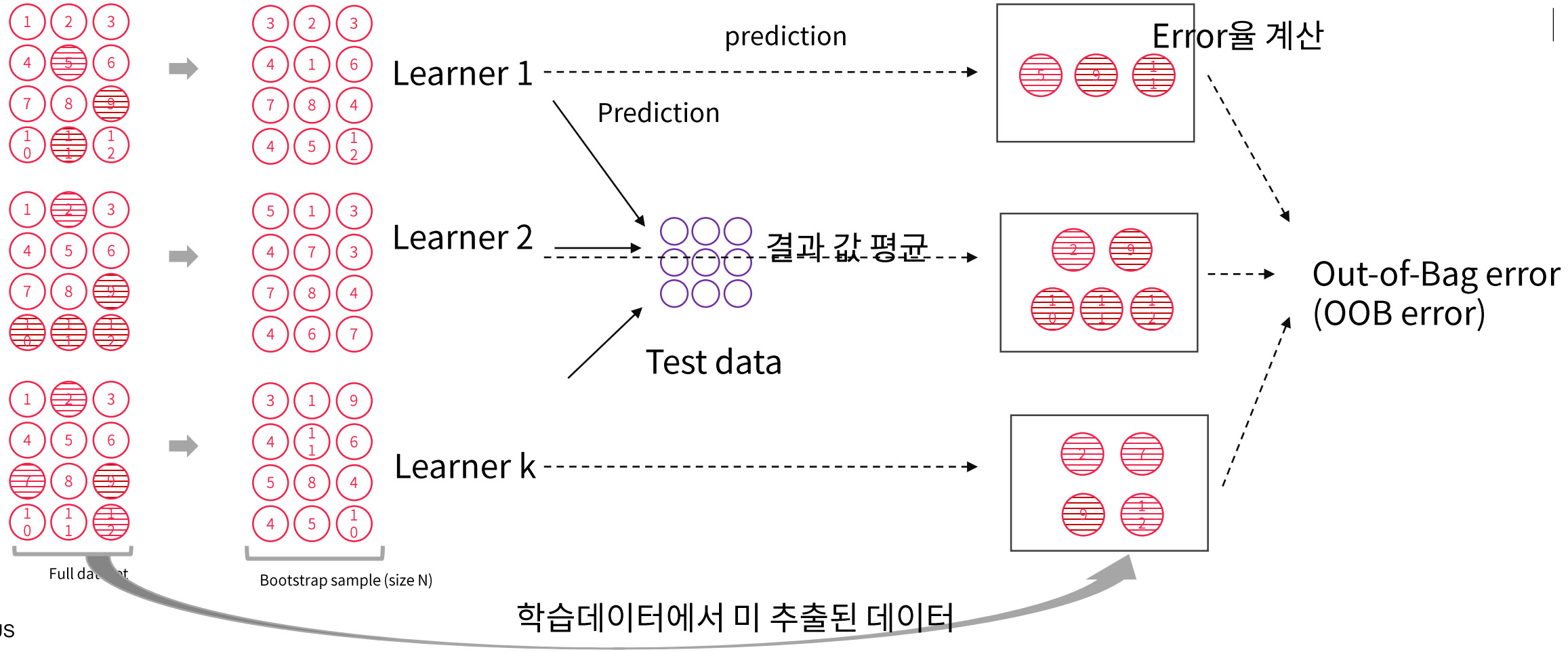
Tree기반의 단일 모델 (패키지 함수)

- Stacking : 모델의 output값을 새로운 독립변수로 사용

“Ensemble의 한 개념”

I Ensemble 기법 review

- **Bagging (bootstrap aggregating)**



I Ensemble 기법 review

- **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

↓

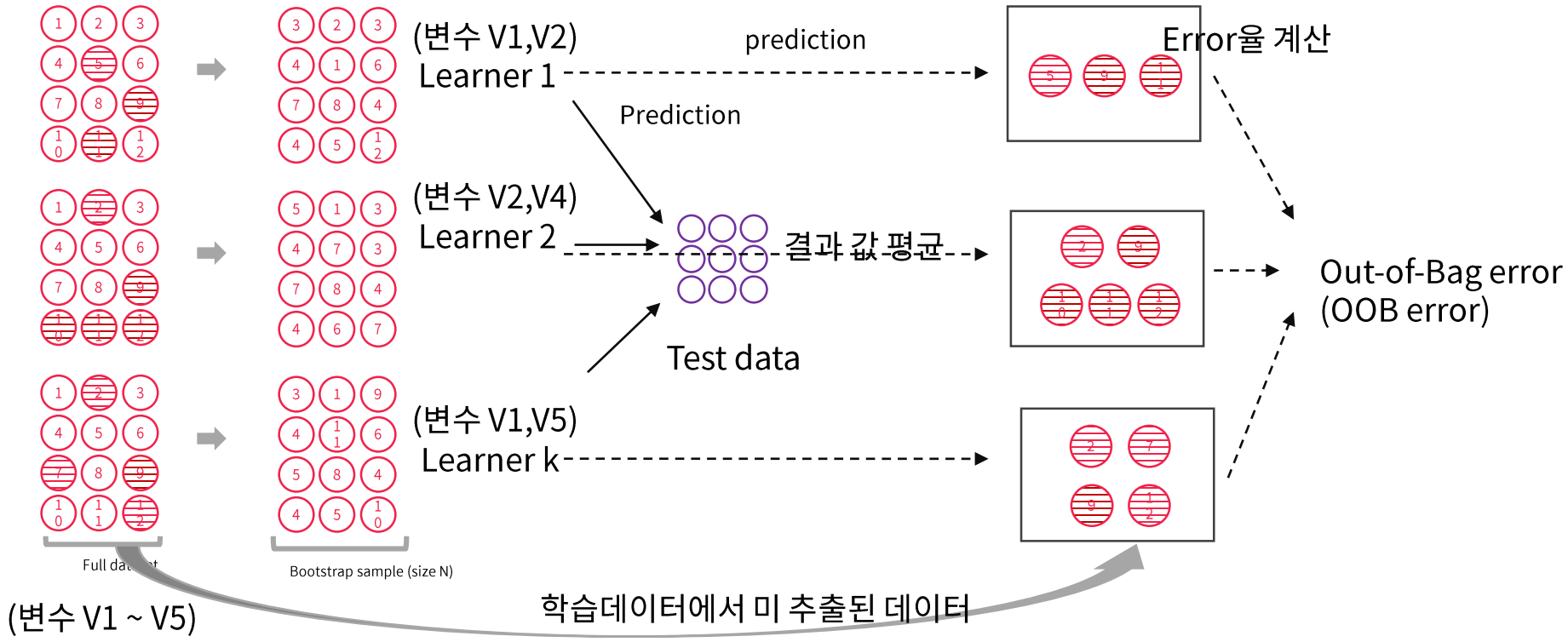
각 Tree간 공분산을 줄일 수 있는 방법이 필요함

I Ensemble 기법 review

- **RandomForest**

  Ensemble learning의 기본 컨셉 - '다양한 모델 '

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



I Ensemble 기법 review

▪ Boosting이란

- Boosting은 **오분류된 데이터에 초점을 맞추어 더 많은 가중치를 주는 방식**
- 초기에는 모든 데이터가 동일한 가중치를 가지지만, 각 round가 종료된 후 가중치와 중요도를 계산
- 복원추출 시에 가중치 분포를 고려
- 오분류된 데이터가 가중치를 더 얻게 됨에 따라 다음 round에서 더 많이 고려됨
- Boosting 기법으로 AdaBoost, LPBoost, TotalBoost, BrownBoost, MadaBoost, LogitBoost, Gradient Boosting 등이 있음

I Ensemble 기법 review

▪ Gradient Boosting

- x를 입력 받아 y를 예측하는 모델 h_0가 있다고 하자.

$$y = h_0(x) + \text{error}$$

- Error가 예측 불가능한 랜덤 노이즈가 아닌 경우, 예측 성능을 올리는 가장 직관적인 방법은 error를 제거하는 것.
- 그렇다면 어떻게 error를 제거할 수 있을까?

$$\text{error} = h_1(x) + \text{error2}$$

$$\text{error2} = h_2(x) + \text{error3}$$

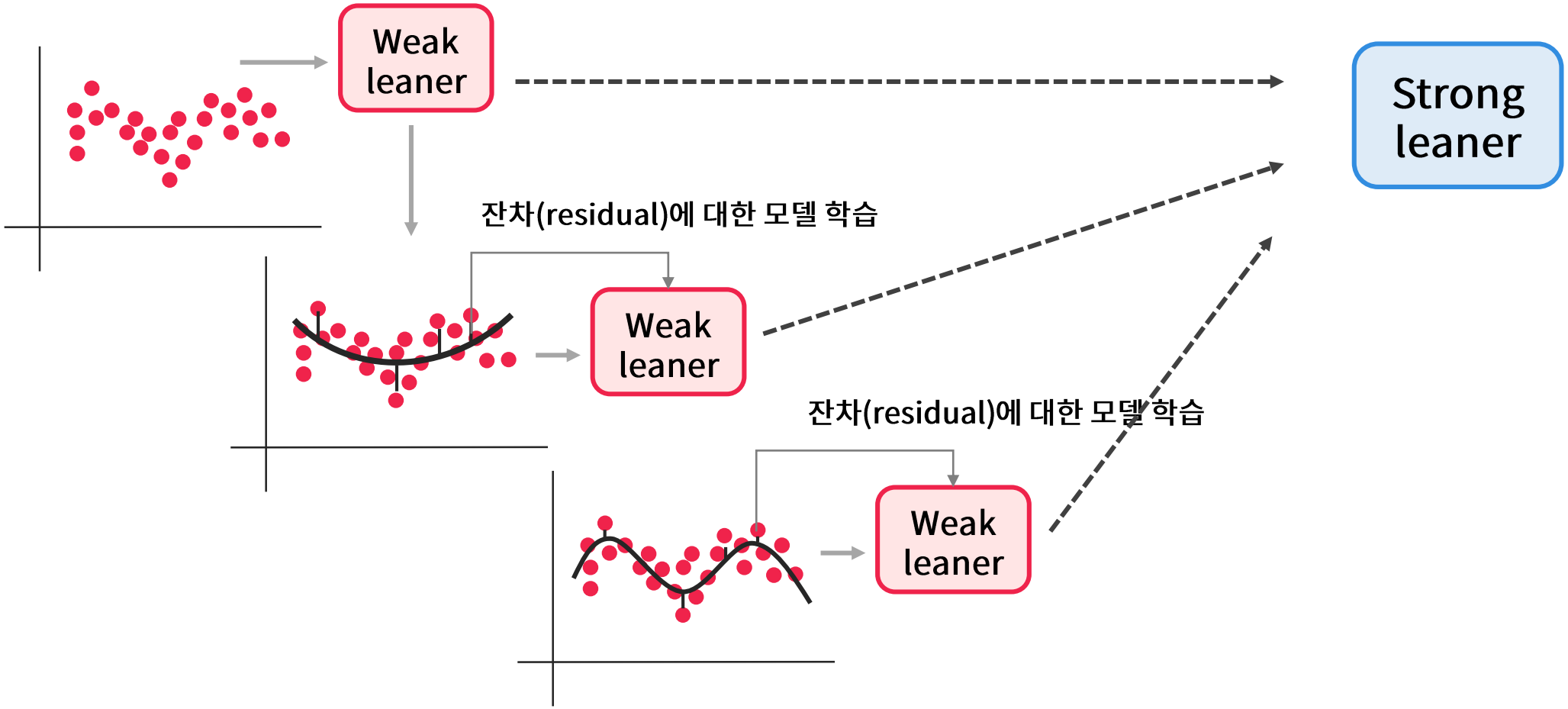
$$\vdots$$

$$\downarrow$$

$$y = h_0(x) + h_1(x) + h_2(x) + \cdots + \text{small error}$$

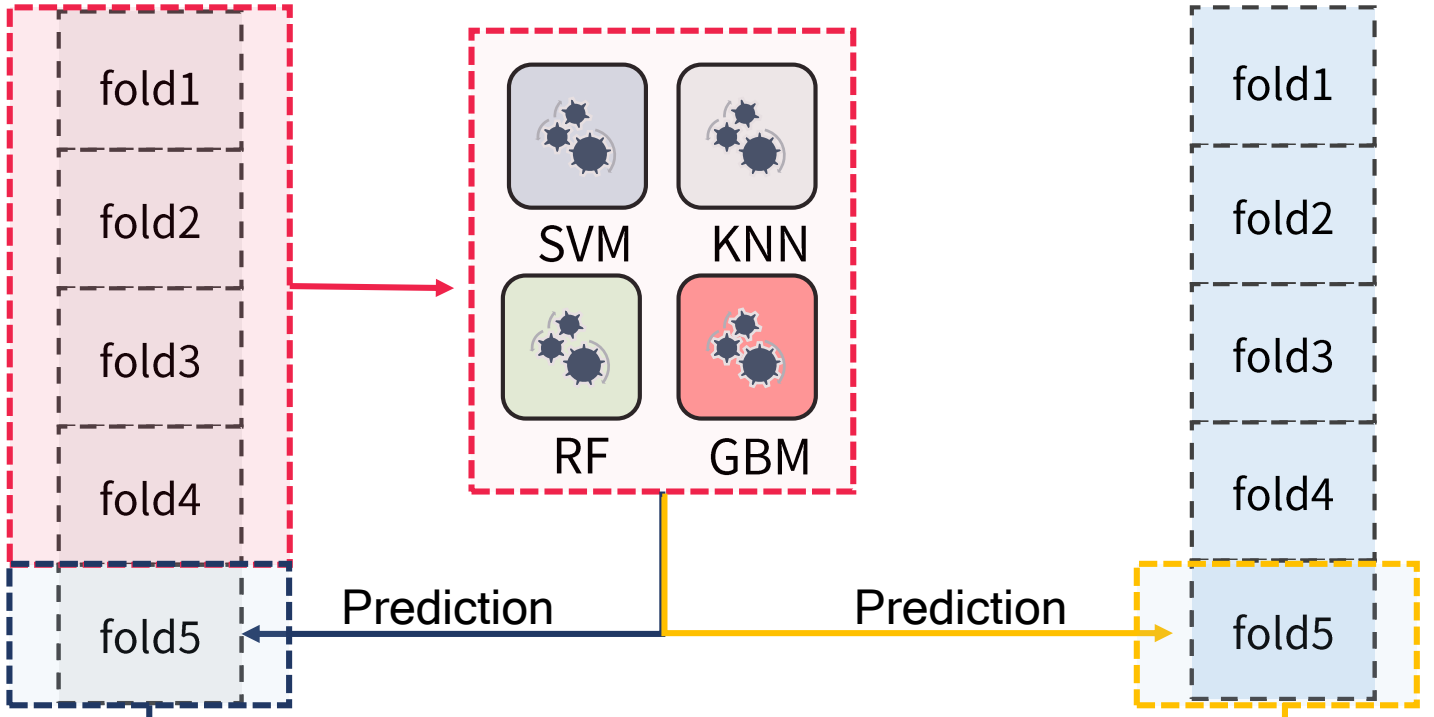
I Ensemble 기법 review

- Gradient Boosting



I Ensemble 기법 review

- Stacking이란



V1	V2
0.2	0.2
0.45	-0.1
-0.5	0.17

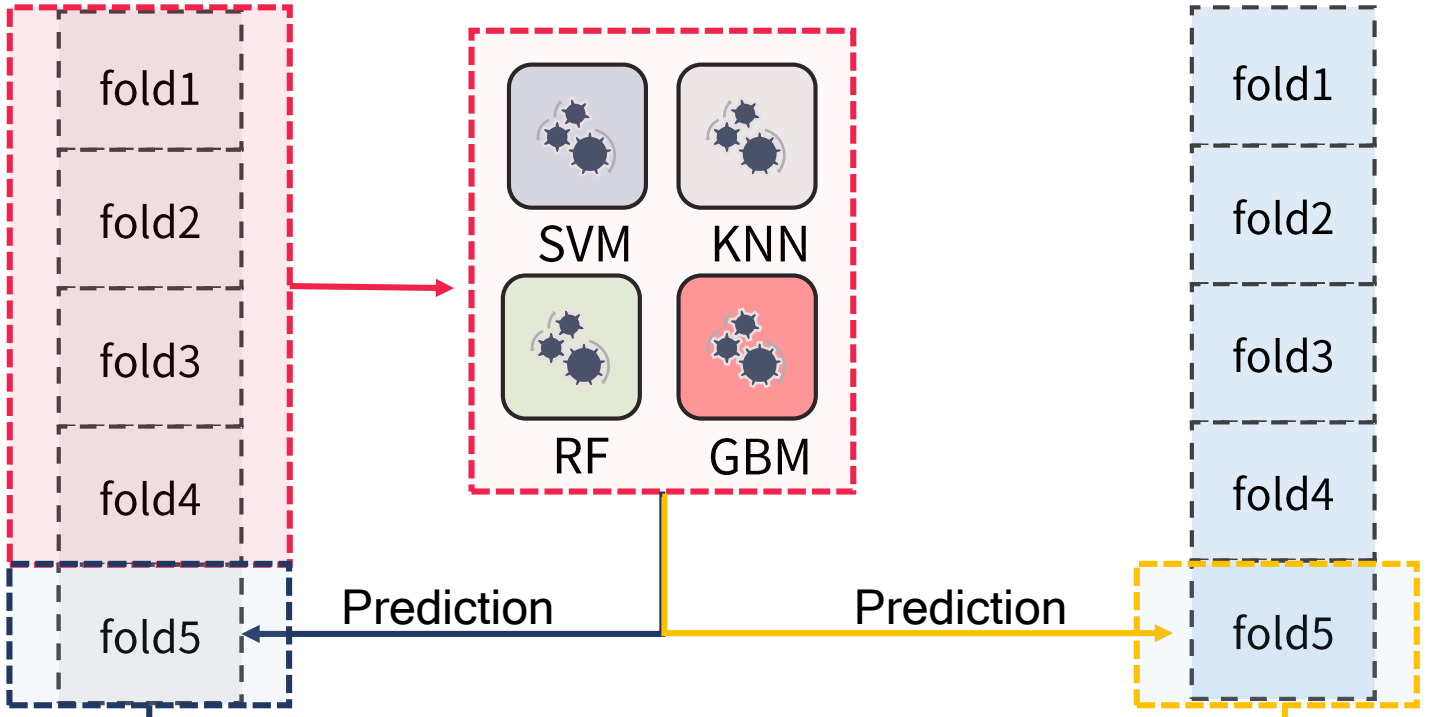
SVM_pr	KNN_pr	RF_pr	GBM_pr
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	불량	불량

V1	V2
0.3	0.42
0.4	-0.01
-0.45	0.17

SVM_pr	KNN_pr	RF_pr	GBM_pr
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	불량	불량

I Ensemble 기법 review

- Stacking이란



V1	V2
0.2	0.2
0.45	-0.1
-0.5	0.17

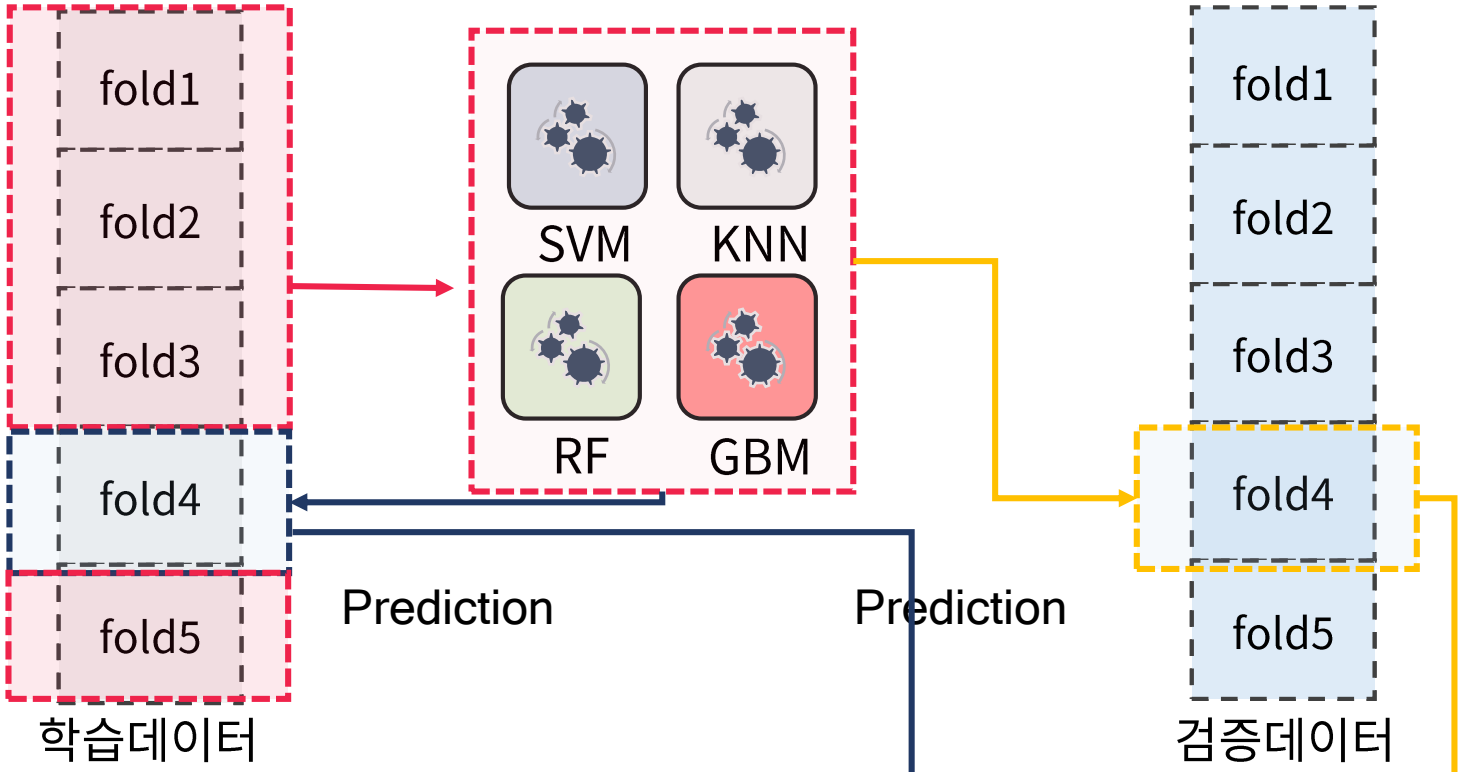
SVM_pr	KNN_pr	RF_pr	GBM_pr
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	불량	불량

V1	V2
0.3	0.42
0.4	-0.01
-0.45	0.17

SVM_pr	KNN_pr	RF_pr	GBM_pr
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	불량	불량

I Ensemble 기법 review

- Stacking이란



V1	V2
0.2	0.2
0.45	-0.1
-0.5	0.17

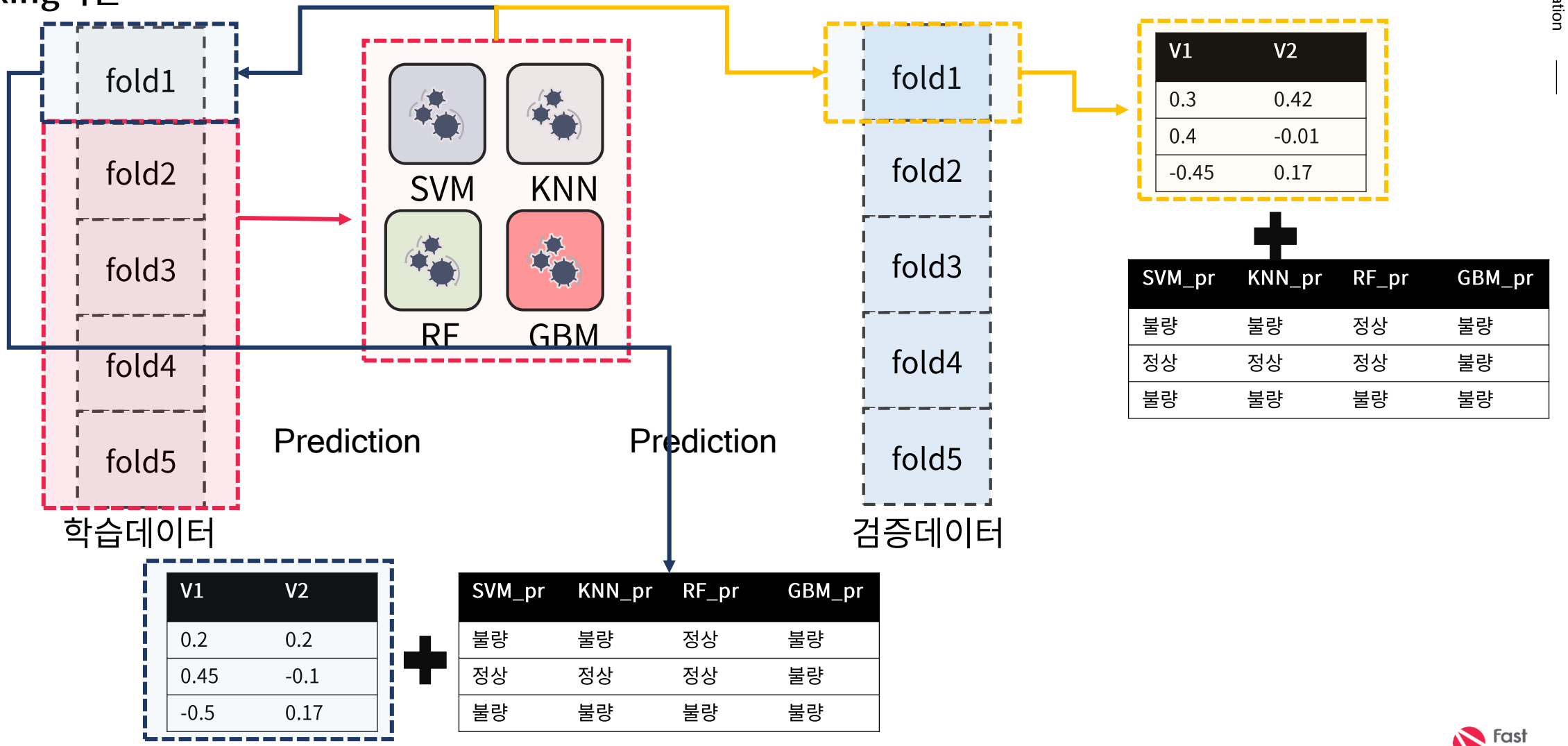
SVM_pr	KNN_pr	RF_pr	GBM_pr
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	불량	불량

V1	V2
0.3	0.42
0.4	-0.01
-0.45	0.17

SVM_pr	KNN_pr	RF_pr	GBM_pr
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	불량	불량

I Ensemble 기법 review

- Stacking이란



V1	V2
0.2	0.2
0.45	-0.1
-0.5	0.17

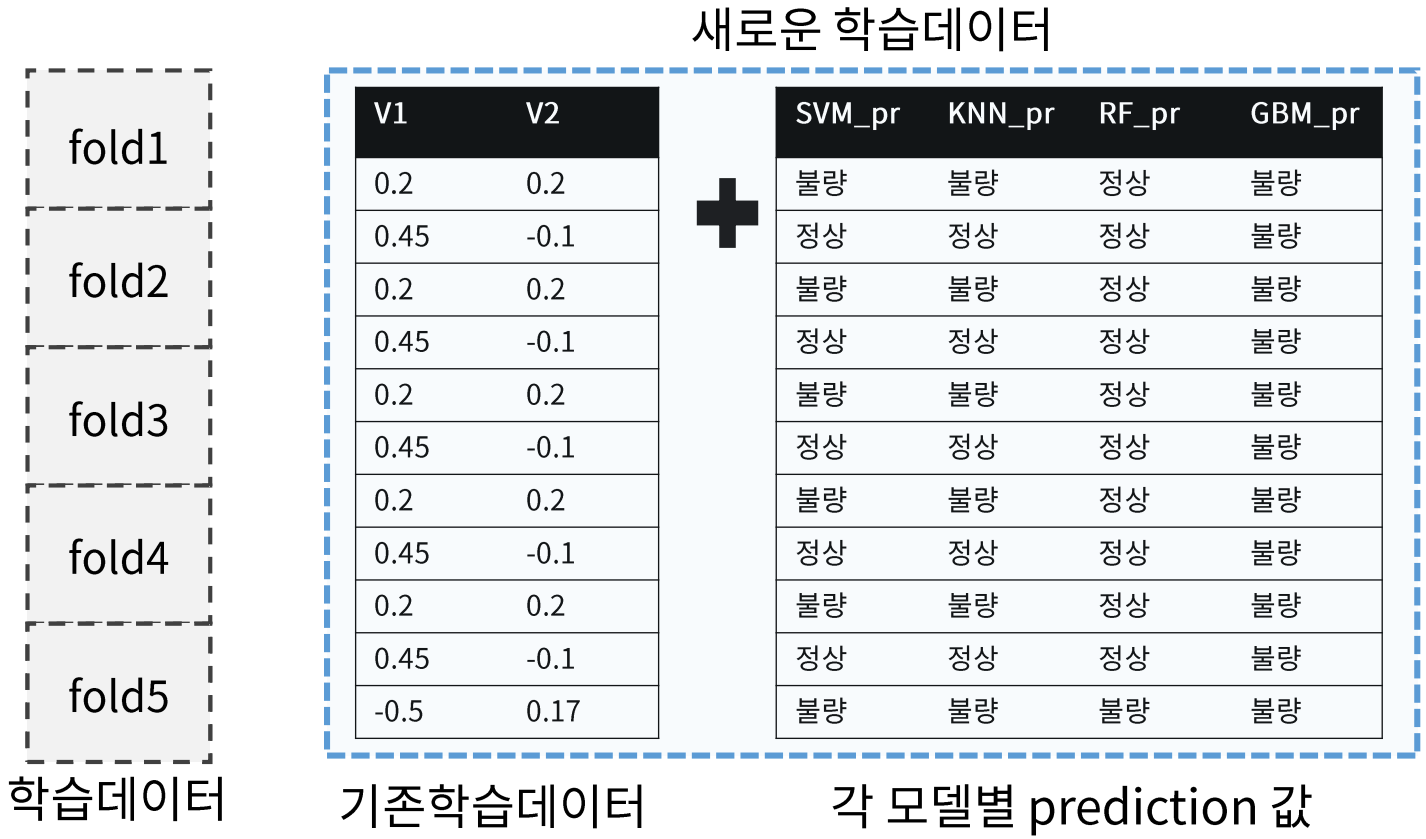
SVM_pr	KNN_pr	RF_pr	GBM_pr
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	불량	불량

V1	V2
0.3	0.42
0.4	-0.01
-0.45	0.17

SVM_pr	KNN_pr	RF_pr	GBM_pr
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	불량	불량

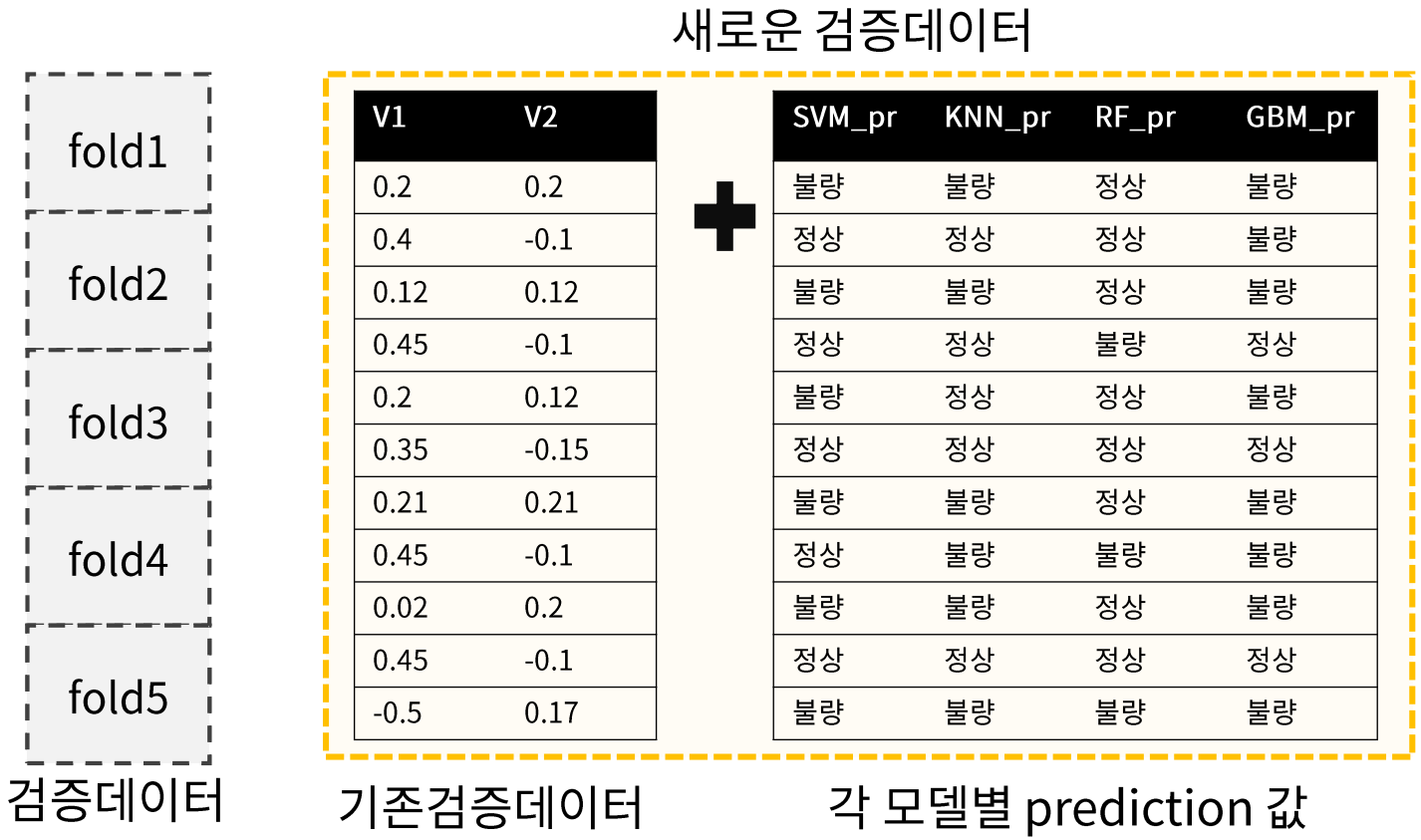
I Ensemble 기법 review

- Stacking이란



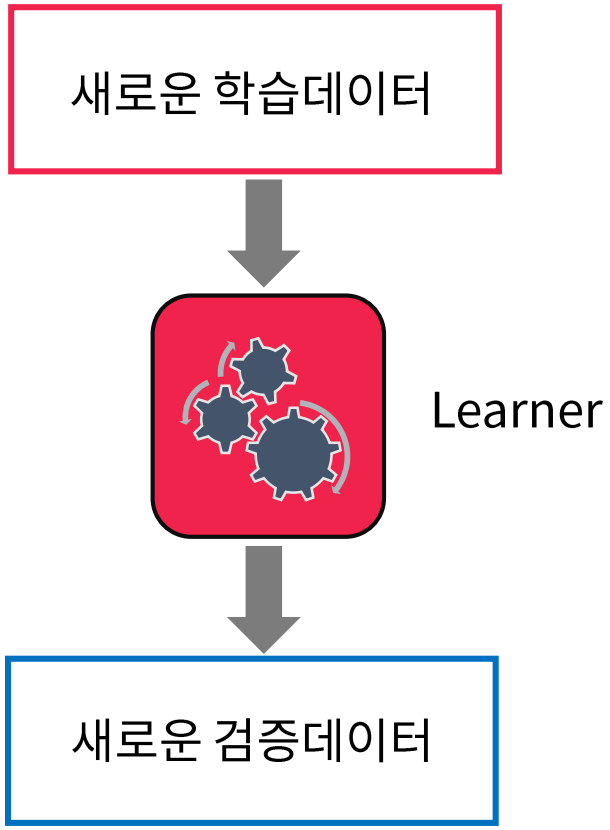
I Ensemble 기법 review

- Stacking이란



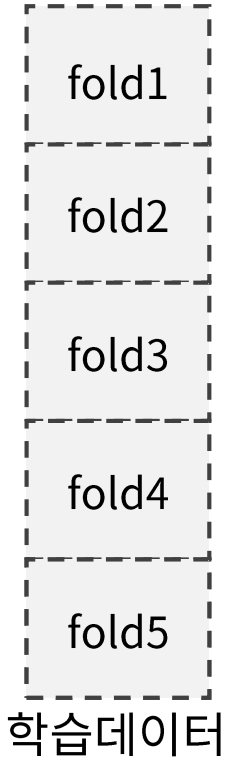
I Ensemble 기법 review

- Stacking이란



I Ensemble 기법 review

- **Stacking이란**
  - 기존 feature를 쓰지 않고 각 모델 별 prediction만을 사용하기도함

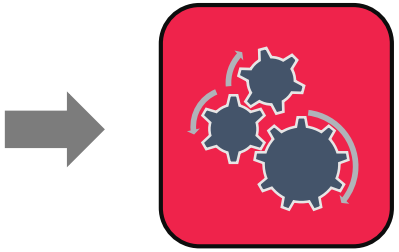


학습데이터

새로운 학습데이터

SVM_pr	KNN_pr	RF_pr	GBM_pr
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	정상	불량
정상	정상	정상	불량
불량	불량	불량	불량

각 모델별 prediction 값

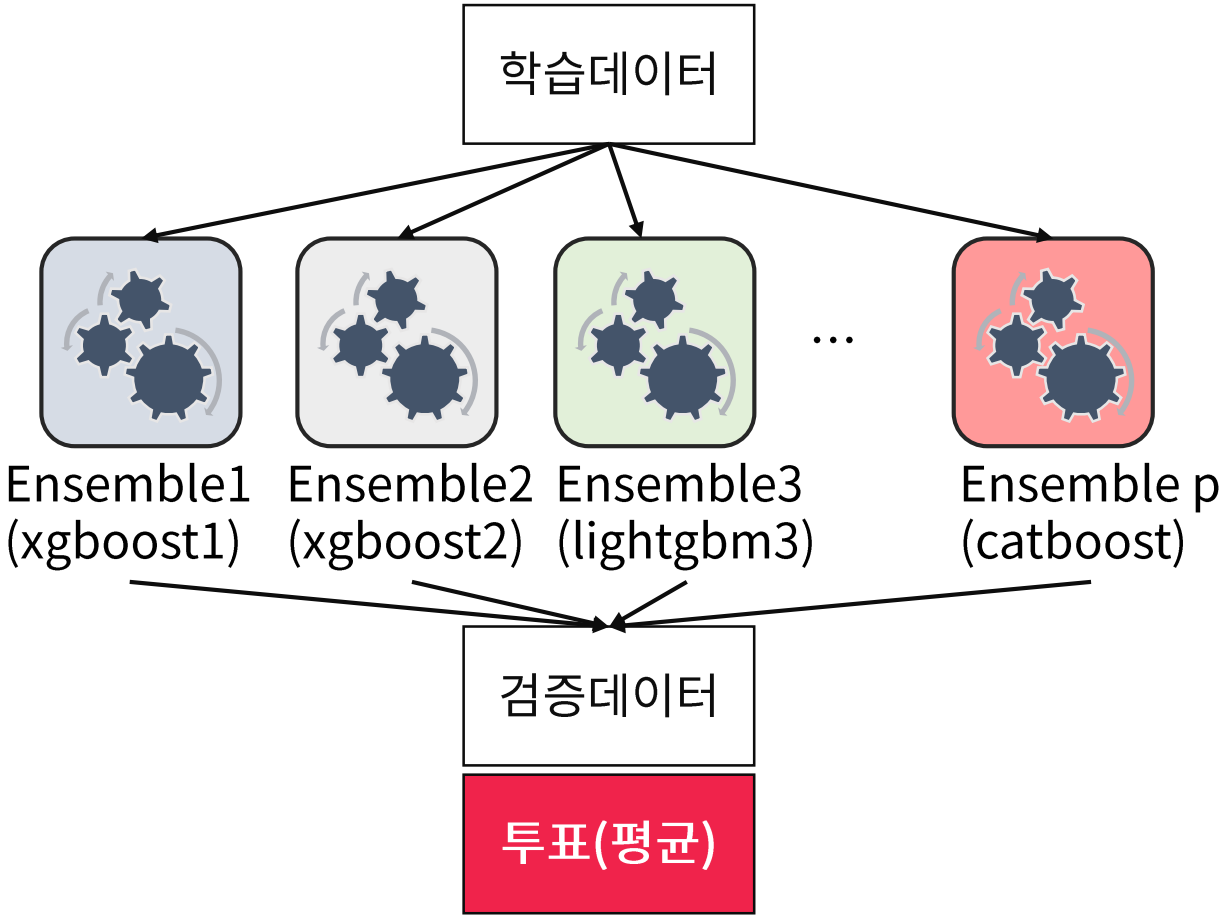


Learner

이 경우, 일반적으로, regression 모델 사용

I Ensemble 기법 review

- **Ensemble의 Ensemble**
  - Ensemble 모델을 단일 모델로 사용하자



Part. 04

Ensemble Learning

중요변수 추출 방법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택