고려대학교 의료정보학과

KU MEDICINE

IMDL

Intelligent Medical Data Lab

의료인공지능

머신러닝 - 부스팅, 스태킹

고려대학교 의료빅데이터연구소
채민수(minsuchae@korea.ac.kr)

1. 부스팅이란

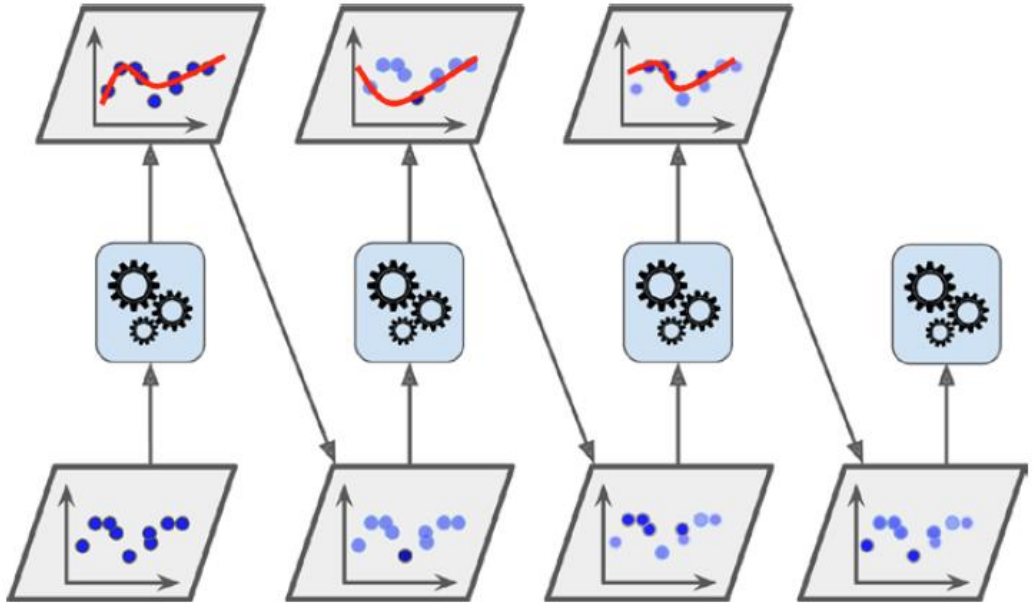
◦ 부스팅(Boosting)

- 가설 부스팅(hypothesis boosting)이란 여러 모델들을 연결하여 강한 예측 모델을 만드는 앙상블 방법
  - AdaBoost(adaptive boosting)
  - Gradient boosting
  - XGBoost(extreme gradient boosting)
  - LightGBM

1. 부스팅이란

◦ AdaBoost

- 잘못 분류된 샘플의 가중치를 높여서 정확도를 개선하고자 하는 아이디어



- Scikit-learn에서는 AdaBoostClassifier와 AdaBoostRegressior를 지원

1. 부스팅이란

◦ AdaBoost - 가중치 계산

- 1) 샘플 가중치 초기화 $w^i = \frac{1}{m}$
- 2) 잘못 분류된 샘플 가중치 적용

$$r_j = \frac{\sum_{\hat{y}^i \neq y^i}^{m} w^i}{\sum_{i=1}^{m} w^i}$$

$$a_j = \eta \log \frac{1-r_j}{r_j}$$

$$w^i = \begin{cases} w^i & \hat{y}^i = y^i \\ w^i \exp(a_j) & \hat{y}^i \neq y^i \end{cases}$$

1. 부스팅이란

◦ AdaBoost - 가중치 계산

- 3) 예측

$$\mathrm{y`}(x) = \underset{k}{\operatorname{argmax}} \sum_{\substack{i=1 \\ \mathrm{y`}(x)=k}}^{n} a_j$$

1. 부스팅이란

- 물고기 데이터셋을 활용하여 AdaBoost 실습
  - 데이터셋 로드
  - 데이터 전처리
  - 입력 특징과 타겟 구분
  - 훈련 데이터와 테스트 데이터 나누기
  - AdaBoost를 활용한 실습

1. 부스팅이란

- 보험료 예측 데이터셋을 활용하여 AdaBoost 실습
  - 데이터셋 로드
  - 데이터 전처리
  - 입력 특징과 타겟 구분
  - 훈련 데이터와 테스트 데이터 나누기
  - AdaBoost를 활용한 실습

1. 부스팅이란

◦ Gradient boosting

- 이전 예측 모델이 만든 잔여 오차를 학습하여 오차를 줄이는 방식

```
previous = target

for i = 0 to m
    clf[i].fit(feature, previous)
    previous = target - clf[i].predict(feature)

y_pred = sum(c.predict(new) for c in clf)
```

1. 부스팅이란

◦ 물고기 데이터셋을 활용하여 Gradient boosting 실습

- 데이터셋 로드
- 데이터 전처리
- 입력 특징과 타겟 구분
- 훈련 데이터와 테스트 데이터 나누기
- Gradient boosting를 활용한 실습

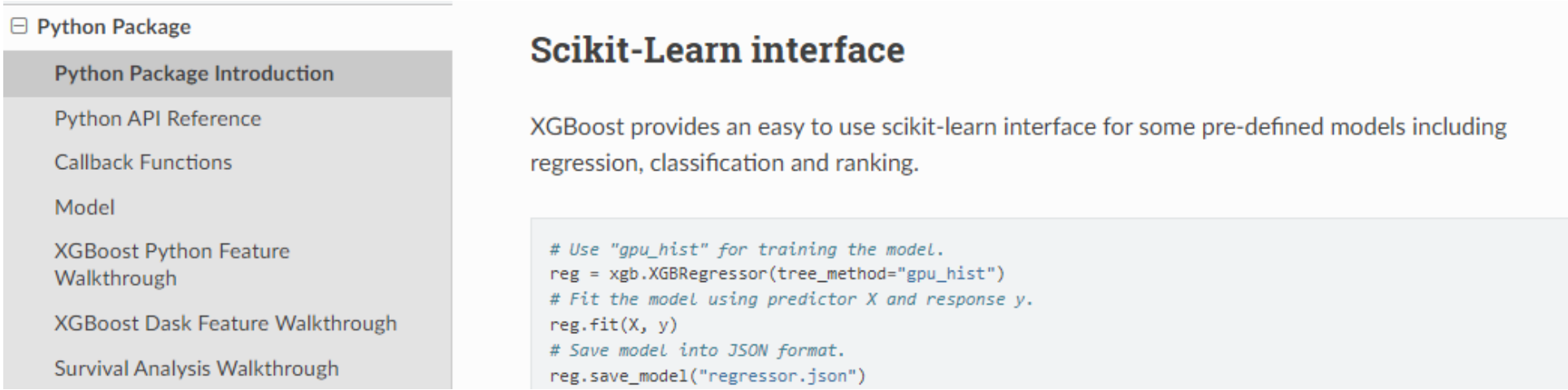
1. 부스팅이란

◦ 보험료 예측 데이터셋을 활용하여 Gradient boosting 실습

- 데이터셋 로드
- 데이터 전처리
- 입력 특징과 타겟 구분
- 훈련 데이터와 테스트 데이터 나누기
- Gradient boosting를 활용한 실습

1. 부스팅이란

- XGBoost
  - 그레디언트 부스팅의 빠른 속도, 확장성, 이식성을 고려하여 개발됨
  - 멀티 스레드를 지원하며, 최근 GPU를 통한 가속 지원
  - https://xgboost.readthedocs.io/



1. 부스팅이란

◦ 물고기 데이터셋을 활용하여 XGBoost 실습

- 데이터셋 로드
- 데이터 전처리
- 입력 특징과 타겟 구분
- 훈련 데이터와 테스트 데이터 나누기
- XGBoost를 활용한 실습

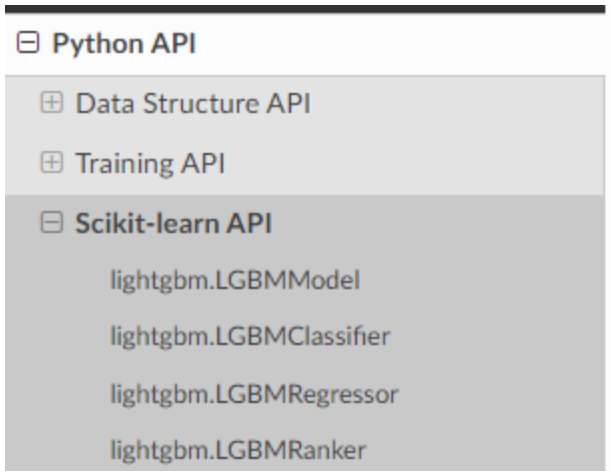
1. 부스팅이란

◦ 보험료 예측 데이터셋을 활용하여 XGBoost 실습

- 데이터셋 로드
- 데이터 전처리
- 입력 특징과 타겟 구분
- 훈련 데이터와 테스트 데이터 나누기
- XGBoost를 활용한 실습

1. 부스팅이란

- LightGBM
  - 빠르면서도 적은 메모리 사용을 위한 그레디언트 부스팅
  - 히스토그램을 통해 빠르면서도 적은 메모리를 사용하도록 함
  - GPU를 통한 가속 지원
  - https://lightgbm.readthedocs.io



Scikit-learn API

`LGBMModel` ([boosting_type, num_leaves, ...])	Implementation of the scikit-learn API for LightGBM.
`LGBMClassifier` ([boosting_type, num_leaves, ...])	LightGBM classifier.
`LGBMRegressor` ([boosting_type, num_leaves, ...])	LightGBM regressor.
`LGBMRanker` ([boosting_type, num_leaves, ...])	LightGBM ranker.

1. 부스팅이란

◦ 물고기 데이터셋을 활용하여 LightGBM 실습

- 데이터셋 로드
- 데이터 전처리
- 입력 특징과 타겟 구분
- 훈련 데이터와 테스트 데이터 나누기
- LightGBM를 활용한 실습

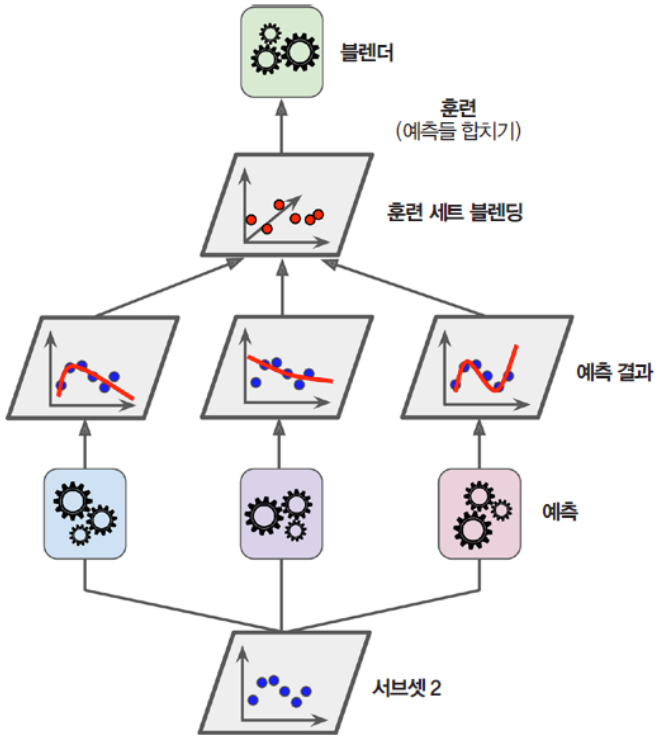
1. 부스팅이란

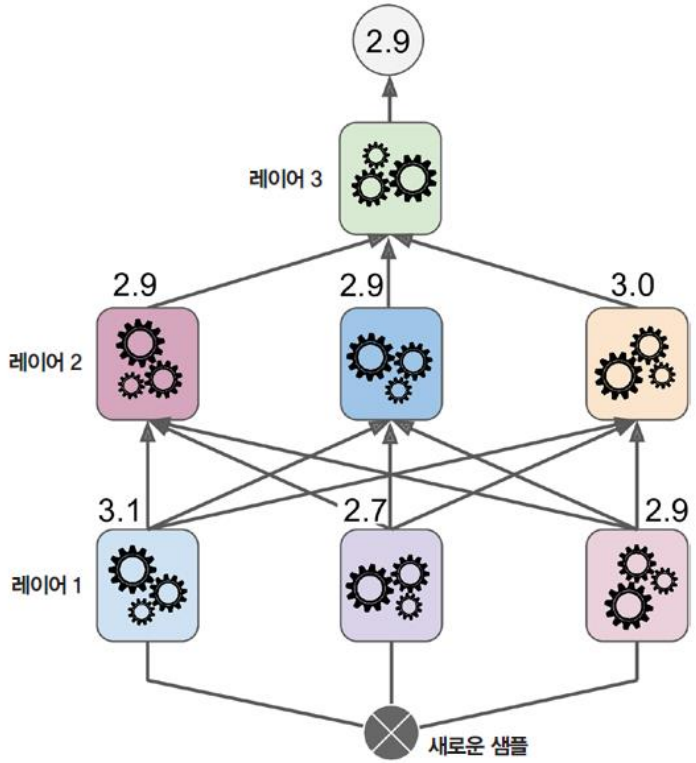
- 보험료 예측 데이터셋을 활용하여 LightGBM 실습
  - 데이터셋 로드
  - 데이터 전처리
  - 입력 특징과 타겟 구분
  - 훈련 데이터와 테스트 데이터 나누기
  - LightGBM를 활용한 실습

2. 스태킹이란

◦ 스태킹(Stacking)

- 앙상블 모델에 속한 예측기를 효과적으로 취합하기 위한 방안





2. 스태킹이란

◦ 물고기 데이터셋을 활용하여 Stacking 실습

- 데이터셋 로드
- 데이터 전처리
- 입력 특징과 타겟 구분
- 훈련 데이터와 테스트 데이터 나누기
- StackingClassifier와 직접 구현한 방식을 통한 실습

2. 스태킹이란

◦ 보험료 예측 데이터셋을 활용하여 Stacking 실습

- 데이터셋 로드
- 데이터 전처리
- 입력 특징과 타겟 구분
- 훈련 데이터와 테스트 데이터 나누기
- StackingRegressor를 활용한 실습

3. 실습 - 1

- 피마 인디언 데이터셋을 이용한 AdaBoost와 Gradient boosting 실습
  - 데이터셋 로드
  - 데이터 전처리
  - 입력 특징과 타겟 구분
  - 훈련 데이터와 테스트 데이터 나누기
  - AdaBoostClassifier와 GradientBoostingClassifier통한 실습

3. 실습 - 2

◦ 심혈관질환 데이터셋을 이용한 XGBoost와 LightGBM 실습

- 데이터셋 로드
- 데이터 전처리
- 입력 특징과 타겟 구분
- 훈련 데이터와 테스트 데이터 나누기
- XGBClassifier와 LGBMClassifier통한 실습

3. 실습 - 3

- 심장질환 데이터셋을 이용한 Stacking 실습
  - 데이터셋 로드
  - 데이터 전처리
  - 입력 특징과 타겟 구분
  - 훈련 데이터와 테스트 데이터 나누기
  - Stacking 을 이용한 실습

4. Homework

◦ 스스로 해보기

- 심장질환 데이터셋을 이용하여 Ada Boost와 Gradient Boosting 을 적용하여 예측하여라.

4. Homework

◦ 스스로 해보기

- 심혈관질환 데이터셋을 이용하여 XGBoost와 LightGBM 을 적용하여 예측하여라.

4. Homework

◦ 스스로 해보기

- 피마 인디언 당뇨병 데이터셋을 이용하여 Stacking 을 적용하여 예측하여라.

➢ F1 score를 기준으로 성능에 따라 점수 차등 부여