Part. 04

Ensemble Learning

Boosting

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택

I Boosting

- **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 Boosting

▪ Bootsing이란

- 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 Boosting

▪ AdaBoost (Adaptive Boost)의 예시

① 모든 데이터에 대해 가중치를 동일하게 0.1로 설정

② Round 1에서 빨간색 네모의 데이터가 수집되며 이를 기반으로 분류 기준값인 B1을 설정(B1보다 작은 경우 +, 큰 경우 –로 분류)

③ 데이터 i에 대해 m번째 round에서의 가중치를 업데이트 **(오분류된 데이터에 가중치를 크게, 정분류된 데이터에 가중치를 작게 설정)**

④ 업데이트한 가중치의 확률로 샘플을 재수집

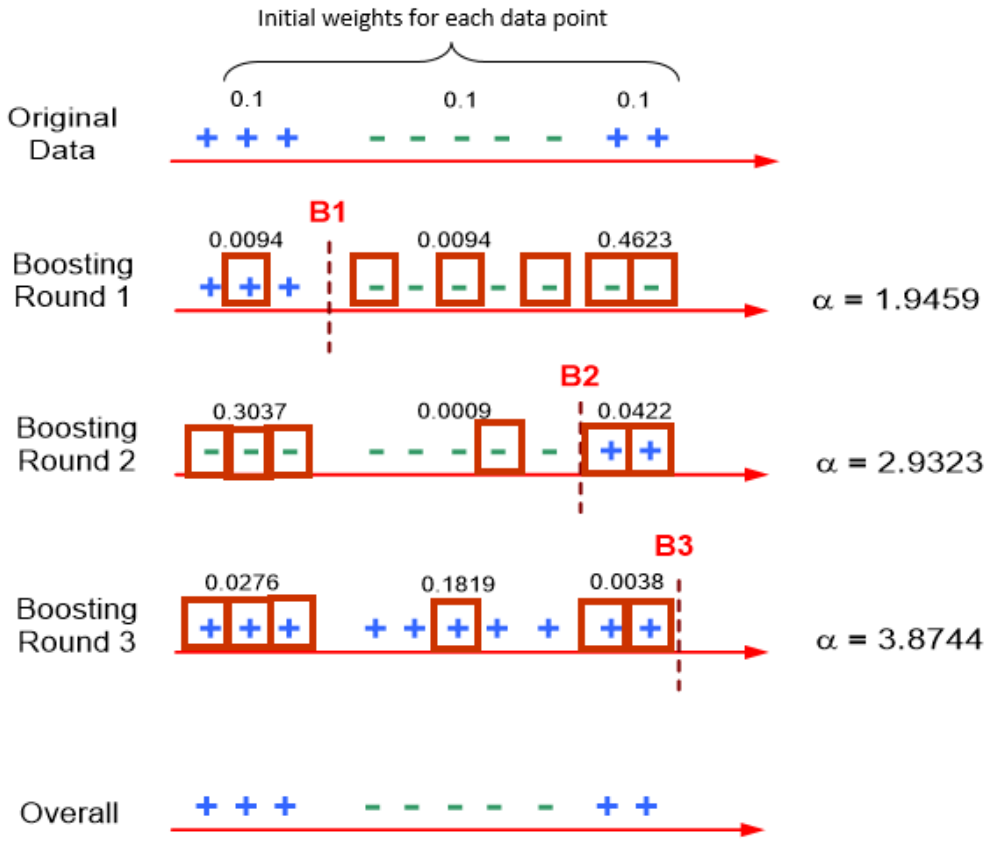
⑤ 4번의 결과로 Round 2의 빨간색 네모의 데이터가 수집

⑥ 수집된 데이터로 모형을 학습한 결과 B2가 분류 기준값으로 도출됨

⑦ 가중치를 업데이트

⑧ 이와 같은 방법을 설정한 반복 횟수만큼 반복

⑨ 예시에서 Round 3까지의 결과를 종합하면 Original Data와 동일한 결과가 도출됨



I Boosting

- **AdaBoost**
  - 최종 의사결정 방법

$$H(x) = sign\left\{\sum_{m=1}^{M} \alpha_m h_m(x)\right\}$$

최종 분류기 (final classifier)

m라운드에서 생성된 약한 분류기 (weak classifier)

m라운드에서 생성된 약한 분류기에 대한 가중치

α_m이 크다는 것은 ε_m이 작다는 의미 (즉 분류기 m이 좋은 성능을 보임)
α_m이 작다는 것은 ε_m이 크다는 의미 (즉 분류기 m이 안 좋은 성능을 보임)

Part. 04

Ensemble Learning

Gradient Boosting

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택