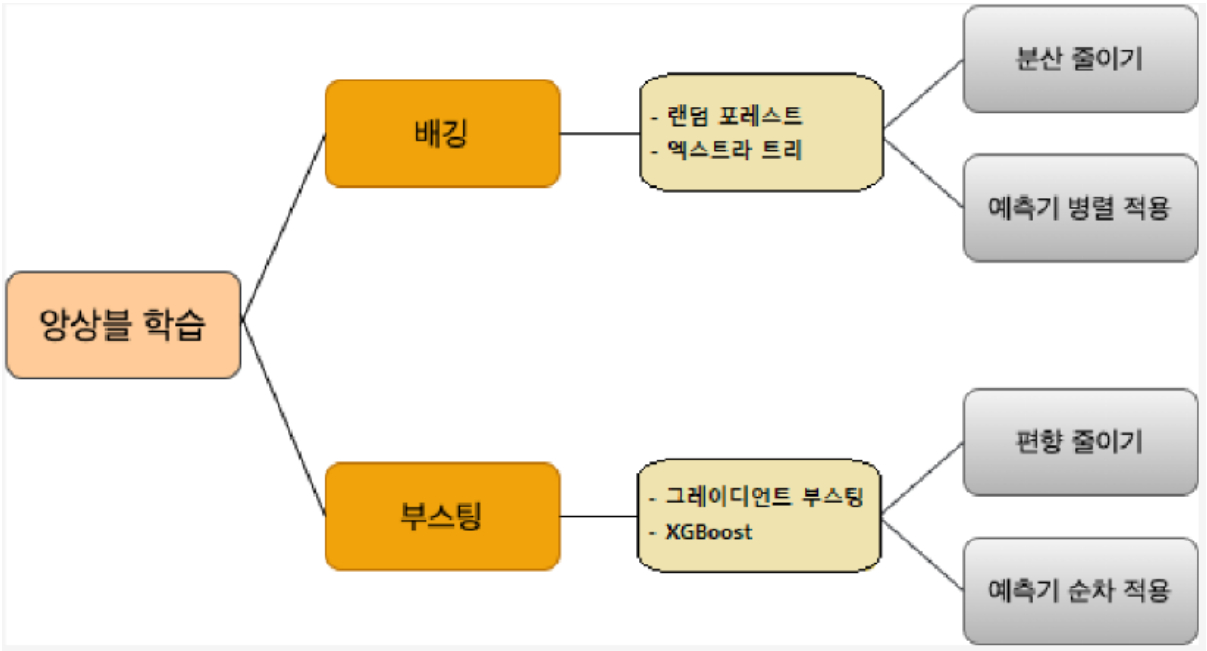
머신러닝 세션 7주차 과제

2024170928 노연경

앙상블 학습 : 여러 개의 모델이 함께 투표해서 최종 결과를 정한다



핵심 아이디어

- 배깅 : 서로 다른 데이터 샘플로 여러 모델 훈련 → 예측 평균/투표

- 부스팅 : 이전 모델이 틀린 샘플에 가중치 높임 → 순차적으로 약한 모델 여러 개 연결

→ 편향과 분산 줄이기

투표식 분류기

: 동일한 훈련셋에 대해 여러 종류의 분류기를 이용하여 앙상블 학습을 적용한 후 직접 또는 간접 투표를 통해 예측값을 결정

- 직접 투표 : 다수결

- 간접 투표 : 예측기들이 예측한 확률의 평균값으로 결정 (높은 확률에 비중을 두기 때문에 직접투표보다 성능이 좋음)

- 앙상블 학습의 성능이 향상되는 이유는 이항분포의 누적분포함수를 이용하여 설명할 수 있음

배깅과 페이스팅

: 여러 개의 동일 모델을 하나의 훈련셋의 다양한 부분집합을 대상으로 학습시키는 방식, 개별 예측기에 비해 분산이 줄어든다. 배깅이 과대적합의 위험을 줄여 배깅 방식이 기본으로 사용된다.

- 배깅 : 중복 허용

- 페이스팅 : 중복 미허용

- OOB 평가 : 배깅 모델에 포함된 예측기로부터 선택되지 않은 훈련 샘플을 이용해 앙상블 학습 모델을 검증하는 기법

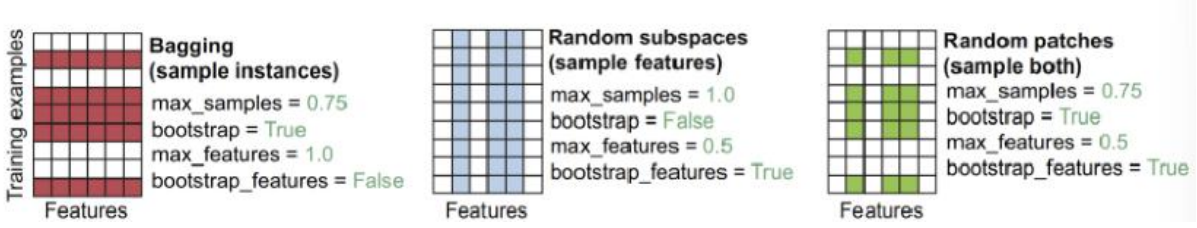
랜덤 패치와 랜덤 서브스페이스

- max_features : 학습에 사용할 특성 수 지정

- bootstrap_features : 학습에 사용할 특성을 선택할 때 중복 허용 여부 지정 (기본값은 False)

- 랜덤 패치 기법 : 훈련 샘플과 훈련 특성 모두를 대상으로 중복 허용, 임의의 샘플 수와 임의의 특성 수 만큼을 샘플링해서 학습하는 기법 (특성, 샘플 다)

- 랜덤 서브스페이스 기법 : 전체 훈련 세트를 학습 대상으로 삼지만 훈련 특성은 임의의 특성 수 만큼 샘플링해서 학습하는 기법 (특성만)



랜덤 포레스트

- 배깅/페이스팅 기법을 적용한 결정트리의 앙상블을 최적화한 모델, 결정트리에 비해 편향은 크고 분산은 낮음

- 특성 중요도 : 불순도를 많이 줄이는 특성은 그만큼 중요도가 커짐, 중요도의 전체 합은 1

부스팅

: 성능이 약한 모델을 순차적으로 보다 강한 성능의 모델로 만들어 가는 기법 (편향을 줄여감)

- 그레디언트 부스팅 : 이전 모델에 의해 생성된 잔차를 보정하도록 새로운 예측기 훈련

- learning_rate을 낮게 정하면 많은 수의 결정트리가 필요하지만 성능은 일반적으로 좋아짐

- 조기종료 : 검증셋에 대해 연속적으로 10번 제대로 개선하지 못하는 경우 훈련 자동 종료

- 확률적 그레디언트 부스팅 : 훈련 데이터의 일부만 무작위로 선택하여 새 모델 학습

- 히스토그램 그레디언트 부스팅 : 대용량 데이터셋을 이용하여 훈련해야 하는 경우 사용, 훈련 샘플의 특성값을 max_bins 개의 구간으로 분류

- XGBoost : extreme gradient boosting

→ 그레디언트 부스팅과의 차이

1. 결정트리 학습에 사용되는 노드 분할을 통해 낮춰야 하는 비용함수가 다름

2. 불순도 대신 모델 훈련의 목적에 맞는 손실 함수 사용 (mse, logloss 등)

3. 생성되는 결정트리의 복잡도가 비용함수에 포함되어 최종적으로 생성되는 모델에 사용되는 결정트리의 복잡도를 가능한 낮추도록 유도