**앙상블 학습과 랜덤 포레스트 요약**

**앙상블 학습 개념**

앙상블 학습은 여러 개의 머신러닝 모델(예측기)을 조합해 단일 모델보다 더 나은 예측 성능을 얻는 방법이다. 대표적으로 배깅, 부스팅, 투표식 분류기 등이 있다. 앙상블은 편향(bias)과 분산(variance) 문제를 완화하는 데 효과적이다.

* 편향: 모델이 정답에 대해 잘못된 가정을 할 때 발생하며, 과소적합의 원인이 된다.
* 분산: 데이터의 작은 변화에 모델이 민감하게 반응하는 현상으로, 과대적합의 원인이 된다.
* 앙상블은 여러 예측기의 결과를 결합해 분산을 줄이고, 때로는 편향도 줄일 수 있다

**투표식 분류기(Voting Classifier)**

* 여러 종류의 분류기를 하나의 앙상블로 묶고, 각 분류기의 예측 결과를 투표로 결정한다.
* 직접투표(hard voting): 다수결로 최종 예측을 결정.
* 간접투표(soft voting): 각 분류기가 예측한 확률의 평균을 내어 가장 높은 확률의 클래스를 선택. soft voting이 보통 더 좋은 성능을 보인다.
* 예측기들이 서로 독립적일수록 앙상블 효과가 크다

**배깅(Bagging)과 페이스팅(Pasting)**

* 배깅: 훈련 데이터에서 중복을 허용하며(부트스트랩) 여러 부분집합을 뽑아 각기 다른 예측기를 학습시키고, 결과를 평균(회귀) 또는 다수결(분류)로 결정한다.
* 페이스팅: 중복 없이 부분집합을 뽑아 예측기를 학습시키는 방식.
* 배깅은 분산을 줄여 과대적합 위험을 낮춘다. 대표적인 배깅 모델이 랜덤 포레스트다

**랜덤 패치와 랜덤 서브스페이스**

* 랜덤 패치: 훈련 샘플과 특성 모두를 무작위로 샘플링해 예측기를 학습.
* 랜덤 서브스페이스: 전체 샘플을 대상으로, 특성만 무작위로 샘플링해 학습.
* 이런 방식은 고차원 데이터에서 다양성을 높여 앙상블 효과를 강화한다

**랜덤 포레스트(Random Forest)**

* 여러 결정트리를 배깅 방식으로 결합한 대표적 앙상블 모델.
* 각 트리는 데이터와 특성의 무작위 샘플을 사용해 학습되고, 분류는 다수결, 회귀는 평균으로 예측한다.
* 랜덤 포레스트는 분산이 낮고, 과대적합 위험이 적으며, 특성 중요도(feature importance)를 제공한다.
* 하이퍼파라미터로는 트리 개수(n\_estimators), 각 트리의 최대 특성 수(max\_features), 최대 리프 노드 수(max\_leaf\_nodes) 등이 있다

**엑스트라 트리(Extra Trees, Extremely Randomized Trees)**

* 랜덤 포레스트와 유사하지만, 트리 분할 시 임계값도 무작위로 선택해 더 빠른 속도를 제공한다.
* 분산은 더 낮아지지만, 편향은 다소 커질 수 있다

**부스팅(Boosting)**

* 여러 약한 예측기를 순차적으로 학습시키며, 이전 예측기의 오류를 보완하는 방식.
* 대표적 부스팅 모델: AdaBoost, Gradient Boosting, XGBoost.
* Gradient Boosting은 이전 모델의 잔차(오차)를 보정하는 새로운 트리를 반복적으로 추가한다.
* 학습률(learning\_rate)과 결정트리 개수(n\_estimators) 조절로 성능과 과적합을 제어한다.
* 조기 종료(early stopping), 부분 샘플링(subsample), 히스토그램 기반 부스팅 등 다양한 규제와 변형 기법이 있다

**앙상블 학습의 실제 활용**

* 표 형식의 정형 데이터 분석에서 뛰어난 성능을 보인다.
* 이미지, 오디오, 자연어 등 비정형 데이터에는 딥러닝이 주로 사용되지만, 딥러닝 모델에도 앙상블 기법을 적용해 성능을 높일 수 있다.
* 캐글(Kaggle) 등 데이터 경진대회에서 상위권 모델은 대부분 앙상블(랜덤 포레스트, XGBoost, 그레디언트 부스팅 등)이다