**7주차 요약: 앙상블 학습과 랜덤 포레스트**

**1. 앙상블 학습(Ensemble Learning)**

* **정의**: 여러 개의 모델을 결합해 더 나은 예측 성능을 얻는 머신러닝 기법.
* **장점**: 편향(Bias)과 분산(Variance)을 동시에 줄여 성능 향상.
* **활용**: Kaggle 대회에서 상위권 모델 대부분이 앙상블 기반(XGBoost, Random Forest 등).

**2. 투표식 분류기(Voting Classifier)**

* **직접 투표 (Hard Voting)**: 각 분류기의 예측값 중 다수결로 최종 예측 결정.
* **간접 투표 (Soft Voting)**: 각 분류기가 예측한 확률의 평균을 기반으로 결정.
* **예시**: 5개 모델이 각각 [1, 1, 3, 3, 3]을 예측한 경우 → Hard Voting은 클래스 3 선택.

간접 투표는 predict\_proba() 기능이 있는 모델에서만 사용 가능.

**3. 배깅(Bagging)과 페이스팅(Pasting)**

* **배깅**: 중복을 허용하는 부트스트래핑 샘플링 사용. → 예측 안정성 향상, 분산 감소.
* **페이스팅**: 중복 없이 샘플링. 배깅과 같은 방식이나 중복만 다름.
* **예측 방식**:
  + 분류: 다수결(hard voting)
  + 회귀: 평균값

예시: RandomForest는 결정트리를 배깅 방식으로 여러 개 조합한 모델.

**4. OOB 평가(Out-of-Bag Evaluation)**

* 배깅 시 각 트리가 일부 샘플만 학습함 → 학습에 사용되지 않은 데이터(oob)를 검증에 사용.
* 별도의 검증셋 없이 모델 평가 가능.
* oob\_score=True 옵션 사용 시 자동 평가.

**5. 랜덤 포레스트(Random Forest)**

* **정의**: 배깅 + 특성 랜덤 선택을 결합한 결정트리 앙상블.
* **특징**:
  + 각 노드 분할 시 전체 특성 중 일부만 무작위로 선택.
  + 편향은 다소 증가하지만 분산은 크게 감소.
  + RandomForestClassifier, RandomForestRegressor 사용.

**6. 엑스트라 트리(Extra Trees)**

* 랜덤포레스트보다 무작위성이 더 강한 모델.
* 노드 분할 시 임계값도 무작위로 선택.
* **장점**: 속도 빠름, 분산 낮음
* **단점**: 편향 증가 가능

**7. 부스팅(Boosting)**

* **정의**: 약한 모델을 순차적으로 학습시켜 성능을 높이는 방식.
* **종류**:
  + **AdaBoost**: 오분류 샘플에 가중치를 증가시켜 다음 모델 훈련.
  + **Gradient Boosting**: 이전 모델의 오차(잔차)를 보정하는 방식.
  + **XGBoost**: 성능 향상 및 속도 최적화된 Gradient Boosting.

**8. 부스팅 관련 하이퍼파라미터**

* learning\_rate: 각 모델의 기여도 조절, 낮출수록 더 많은 트리 필요하지만 일반적으로 성능 향상.
* n\_iter\_no\_change: 일정 횟수 이상 성능 향상이 없으면 학습 조기 종료.
* subsample: 훈련 샘플의 일부만 사용하여 학습. → 훈련 속도 ↑, 분산 ↓

**9. 특성 중요도 (Feature Importance)**

* **의미**: 결정트리가 특정 특성을 얼마나 자주, 얼마나 깊은 위치에서 사용했는지에 기반해 중요도 평가.
* **사용**: feature\_importances\_ 속성에서 확인 가능.
* **예시**: 붓꽃 데이터셋에서 꽃잎 길이 등이 높은 중요도로 나타남.