Ch7. 앙상블 & 랜덤 포레스트

📍앙상블 학습

-여러 개의 모델을 훈련시킨 결과를 이용하는 기법

-편향과 분산을 줄이는 것이 중요함

*편향: 편향이 크면 과소적합 발생

*분산: 분산이 크면 과대적합 발생

*편향과 분산의 트레이드오프: 편향과 분산을 동시에 좋아지게 할 수는 없음

📍앙상블 학습의 종류

-**배깅 기법**: 여러 개의 예측기를 독립적으로 훈련, 예측값들의 평균값을 최종 모델의 예측값으로 활용

-**부스팅 기법**: 여러 개의 예측기를 순차적으로 훈련한 최종 예측값을 사용

📍투표식 분류기

-동일한 훈련셋에 대해 여러 종류의 분류기를 이용하여 앙상블 학습을 적용한 후 직접/간접 투표를 통해 예측값 결정

-**직접 투표**: 예측값들의 다수로 결정

-**간접 투표**: 예측기들의 예측한 확률값들의 평균값으로 예측값 결정

📍배깅과 페이스팅

-여러 개의 동일 모델을 하나의 훈련셋의 다양한 부분집합을 대상으로 학습시킴

-**배깅**: 훈련셋의 부분집합을 선택할 때 중복 허용 샘플링

-**페이스팅**: 중복 미허용 샘플링

***OOB 평가**: 각각의 샘플에 대해 해당 샘플을 훈련에 사용하지 않은 모델들의 예측값을 이용하여 앙상블 학습 모델을 검증하는 기법

📍랜덤 패치와 랜덤 서브스페이스

-BaggingClassifier는 특성에 대한 샘플링 기능을 지원함

-max_features: 학습에 사용할 특성 수 지정(정수-지정된 수만큼 특성 선택, 부동소수점-지정된 비율만큼 특성 선택)

-bootstrap_features: 특성 중복 허용 여부 지정

-랜덤 패치 기법: 훈련 샘플&훈련 특성 모두 중복 허용, 임의의 샘플 수와 임의의 특성 수 만큼을 샘플링해서 학습함

-랜덤 서브스페이스 기법: 훈련 샘플은 전체 대상, 훈련 특성은 임의의 특성 수만 샘플링하여 학습함

📍랜덤 포레스트

-배깅/페이스팅 기법을 적용한 결정트리의 앙상블을 최적화한 모델

-**RandomForestClassifier**: 분류 용도

-**RandomForestRegressor**: 회귀 용도

📍부스팅

-순차적으로 이전 학습기의 결과를 바탕으로 예측값의 정확도를 조금씩 높여가면서 성능이 약한 모델을 순차적으로 보다 강한 성능의 모델로 만들어감

-그래디언트 부스팅: 이전모델에 의해 생성된 잔차를 보정하도록 새로운 예측기 훈련 (GradientBoostingClassifier, GradientBoostingRegressor)

-확률적 그래디언트 부스팅: 훈련 샘플의 특성값을 max_bins 개의 구간으로 분류

-히스토그램 그레디언트 부스팅

-XGBoost