핸즈온 ML CH7 앙상블 & 랜덤포레스트

7.1 투표기반분류기

weak learner 1 ① hard voting_다수결
weak learner 2 → strong learner
weak learner 3 →
weak learner 4 →
:

(큰수의법칙)

② soft voting_개별 클래스예측 확률의 평균으로 결정 (Better)

ex) sklearn Voting Classifier
(voting = "hard" or "soft")
↳ 예측확률 반환 선행필수

SVC: predict_probar or probability = True

7.2 배깅 & 페이스팅 : 같은 알고리즘, 훈련세트 무작위 병렬 수행 가능, 분산↓ 편향 -

↓ 중복O (bootstraping) ↓ 중복X

모든 예측기의 예측 모아서
- 분류: 최빈값
- 회귀: 평균계산

Sklearn Bagging Classifier

배깅: m개 샘플 중복하여 random 선택
약 63%만 샘플링 됨 → 나머지 37% : "oob 샘플"
테스트 데이터로 이용
각 예측기의 oob평가 평균으로 앙상블평가
oob_decision_function에 예측확률 저장됨

7.3 랜덤 패치와 랜덤 서브스페이스 분산↓ 편향↑

특성샘플링 : 무작위로 특성을 선택 bootstrap_features = True, max_features < 1

이미지와 같은 고차원 dataset에서 유용

① 특성, 샘플 모두 샘플링하는 것 → 랜덤 패치방식
② 특성만 샘플링하는 것 → 랜덤서브스페이스

7.4 랜덤포레스트

배깅(or 페이스팅) 적용한 결정트리 앙상블 (RandomForestClassifier / Regressor (Bagging_Classifier + DecisionTree Classifier)

* 익스트림 랜덤트리 (엑스트라트리) 앙상블 분산↓ 편향↑
: 트리에서 매 노드마다 최적의 임계값 찾는 대신 무작위로 분할한 뒤 최상의 분할 선택.
(≒ ExtraTreesClassifier)

* 특성 중요도 측정
각 노드에서 얼마나 불순도 감소되었는가? (=가중치 평균=연관된 훈련샘플수)
feature_importances_ 에 저장됨

7.5 부스팅 boosting

① AdaBoost : 이전 모델에서 라소적합된 샘플의 가중치 높여가며 … / 연속적 알고리즘. 병렬 실행 불가

$$w^{(i)} = \begin{cases} w^{(i)} & \hat{y}_j^{(m)} = y^{(i)} \\ w^{(i)} \exp(\alpha_j) & \hat{y}_j^{(m)} \neq y^{(i)} \end{cases}$$

$$\hat{y}(x) = \underset{k}{\operatorname{argmax}} \sum_{\substack{j=1 \\ \hat{y}_j(x)=k}}^{N} \alpha_j$$

N : 예측기 수

→ 가중치 합이 가장 큰 클래스가 예측결과가 됨

sklearn AdaBoostClassifier

② GradientBoosting : 이전 예측기가 만든 잔여 오차 (residual error)에 새로운 예측기 학습

$y_2 = y - \text{predict}$

sklearn GradientBoostingRegressor (learning_rate)

낮게(0.1) 설정하면 트리가 많이 필요하나 성능↑

∴ 축소규제

+) SGB

XGBoost 이게뭘지..

7.6 스태킹 : 블렌더 or 메타학습기 생성