머신러닝(Machine Learning) 모델링

작성자 : 김진성

머신러닝(Machine Learning) 모델(제2유형)

1. 선형회귀(Linear Regression) 모델(5회 출제)

✓ 종속변수가 연속형(등간/비율)

예) 매출액, 자동차연비, 영화관객수 등

2. 분류(Classification) 모델(2~4회 출제)

✓ 종속변수가 이산형(이항 또는 다항)

예) 생존여부, 날씨유무, 성별, 혈액형, 와인유형 등

머신러닝 모델링 예측력 향상 방안

1. 데이터 전처리

- ✓ 결측치와 이상치 처리

2. 특징 공학(Feature Engineering)

- ✓ 테이블의 컬럼(특징) 선택/제거,추가(파생변수),변형(인코딩, 스케일링)

3. 모형 앙상블

- ✓ 앙상블 모델 : RandomForest, XGBoost 선택

4. 초매개변수 최적화

- ✓ Grid-search 이용 best 파라미터 찾기

머신러닝 모델링 절차

단계1. 데이터 전처리 & 특징공학

- ✓ 결측값, 이상치, 변수선택/제거/추가(파생변수), 스케일링, 인코딩 등

단계2. 훈련셋/검증셋 나누기

- ✓ train_test_split, KFold 등

단계3. 사용할 모델 정하기

1) 알고리즘 선택 : RandomForest, XGBoost, LinearRegression 등
2) 모델학습 : model.fit(훈련셋 X변수, 훈련셋 y변수)
3) 모델검증 : 평가지표(roc_auc_score, accuracy_score, MSE, r2_score 등)

단계4. best 파라미터 찾기 : 시간 부족 시 생략

- ✓ Grid-search -> 모델학습 -> 모델검증

단계5. test 데이터 예측값 구하기

- ✓ predict(테스트셋 X변수) : class 예측 , predict_proba(테스트셋 X변수) : 확률예측

단계6. 예측값 csv 파일 작성 & 제출

- ✓ 예측값.to_csv('파일명', index = False)

● RandomForest 하이퍼라미터

파라미터 명	설명
n_estimators	- 결정트리의 갯수를 지정 - Default = 100 - 무작정 트리 갯수를 늘리면 성능 좋아지는 것 대비 시간이 걸릴 수 있음
min_samples_split	- 노드를 분할하기 위한 최소한의 샘플 데이터수 → 과적합을 제어하는데 사용 - Default = 2 → 작게 설정할 수록 분할 노드가 많아져 과적합 가능성 증가
min_samples_leaf	- 리프노드가 되기 위해 필요한 최소한의 샘플 데이터수 - min_samples_split과 함께 과적합 제어 용도 - 불균형 데이터의 경우 특정 클래스의 데이터가 극도로 작을 수 있으므로 작게 설정 필요
max_features	- 최적의 분할을 위해 고려할 최대 feature 개수 - Default = 'auto' (결정트리에서는 default가 none이었음) - int형으로 지정 →피처 갯수 / float형으로 지정 →비중 - sqrt 또는 auto : 전체 피처 중 √(피처개수) 만큼 선정 - log : 전체 피처 중 log2(전체 피처 개수) 만큼 선정
max_depth	- 트리의 최대 깊이 - default = None → 완벽하게 클래스 값이 결정될 때 까지 분할 또는 데이터 개수가 min_samples_split보다 작아질 때까지 분할 - 깊이가 깊어지면 과적합될 수 있으므로 적절히 제어 필요
max_leaf_nodes	리프노드의 최대 개수