HW4, Competition using Ensemble - Big Data

20170735
컴퓨터공학과
이승준

1. Binary Classification using Ensemble

먼저, 이용한 ensemble model은 sklearn.ensemble.GradientBoostingClassifier 입니다.
링크 : https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

이 모델은 수업 시간에 배우지는 않았지만, ensemble 모델이고, 구글링 결과 Gradient Boosting 알고리즘은 기본적으로 Decision tree 를 사용합니다. 그래서 제약 조건에서 벗어나지 않기에, 이 모델을 이용했습니다.

HW3와 거의 비슷하게 전처리 했습니다. 각 Categorical string data feature 들에 대해서, y에 가장 큰 영향을 주는 것부터 오름차순으로 0 부터 assign 했습니다. 그리고, 'balance', 'duration' 등의 numerical data들은 seaborn library의 box plot 메서드를 이용해 outlier 들을 파악하고,
df2.loc[df2['balance']>=35000,'balance']=35000 과 같은 코드로 outlier들을 조정해줬습니다.

그리고 이 gradient boosting classifier 을 선택한 것은, All_discrete_emsemble.py 에서 주석처리 된 코드들로, 테스트 해 본 결과입니다. 각각의 learning rate 들과 n_estimators 들을 바꿔가며 GradientBoostingClassifier, AdaBoost, RandomForest 들을 테스트 해봤습니다. 테스트에는 model_selection.train_test_split 을 이용해 트레이닝 데이터들을 잘라서 사용했습니다.

이에 gradient boosting classifier 를 선택하고, 최적의 learning rate, n_estimator 를 찾아 y 를 도출했습니다.

Public Score : 0.54502

2. Crime Category Classification using ensemble

1번과 거의 비슷하게 preprocessing 했습니다. 그런데, 날짜와 시간을 컬럼을 나눴고, 년, 월, 일에 각각의 가중치를 둬서 Dates 컬럼을 만들었고, 시간과 분에 각각 가중치를 줘서 시간 컬럼을 만들었습니다. 그리고 1번과 같이 여러 번의 반복으로 최적의 learning rate 와 n_estimator 를 찾아서 prediction 했습니다. 이 문제에서도 GradientBoostingClassifier 를 사용했습니다.

Public Score : 2.37098