HW2_34_A #2 (4조)

1. 데이터 요약

- 데이터 형태: 284,807 row와 31개의 column으로 이루어진 테이블
- Column은 시간 정보를 포함하고 있으며, 28개의 이름을 알 수 없는 column과 amount column, 그리고 class column으로 구성
- 테이블에는 null value를 가지고 있지 않음
- Class가 1인 경우는 0.17% 수준으로 매우 적은 비율을 가지고 있는 불균형 데이터
- 우리가 예측하고자 하는 것은 Time과 Amount를 포함한 모든 column 정보를 이용하여 Class를 분류하는 것임

2. 데이터 전처리

- 위 데이터는 Class 1인 경우가 전체의 0.17% 수준으로 매우 불균형함
- 현재 데이터로 랜덤하게 추출하여 training 후 예측할 때에서 여전히 데이터 분포는 불균형한 상태이기 때문에 예측 결과에 왜곡 발생 가능성 있음
- 따라서 test data는 class 0과 class 1의 개수를 각각 100씩 동일하게 구성하여 머신러닝 알고리즘에 따른 성능을 평가함
- 총 6개의 머신러닝 알고리즘 적용함

(Logistic regression, support vector regression, k-nearest neighborhood, neural network, decision tree, Bayesian classification)

- Time, Amount 등과 같이 불필요한 컬럼 제거 전후의 결과 비교

3. 알고리즘별 분석 결과

<u>사용한 머신러닝 알고리즘</u>

(1) Logistic regression

- sklearn.linear_model module의 LogisticRegression() 이용하여 default 조건에서 분석

(2) support vector classification

- sklearn.svm module의 SVC() 이용하여 gamma='scale' 및 나머지는 default 조건에서 분석

(3) K-nearest neighborhood

- sklearn.neighbors module의 KNeighborsClassifier() 이용하여 default 조건에서 분석

(4) Neural network

- sklearn.neural_network module의 MLPClassifier() 이용하여 learning_rate_init=0.01 조건에서 분석

(5) Decision tree

- sklearn.tree module의 DecisionTreeClassifier() 이용하여 분석

(6) Bayesian classification

- sklearn.naive_bayes module의 BernoulliNB() 이용하여 분석

불필요한 컬럼 제거

- 시각화 과제에서 분석한 결과를 바탕으로 class에 영향을 주지 않는 몇몇 컬럼을 제거하고 분석시 결과에 어떤 영향을 주는지 분석함
- 예를들어 Time, Amount 컬럼의 경우 class와 상관없는 컬럼임
- 불필요한 컬럼 제거시 전반적으로 개선될 결과를 보임

분석 조건 및 결과

1) 모든 컬럼 사용: X = ['Time' : 'Amount'], y = ['Class']
2) Time, Amount 컬럼 제거: X = ['V1: 'V28'], y = ['Class']
3) V13, V15, V20 컬럼 추가 제거:

알고리즘		Logistic	SVC	KNN	NN	DT	Bayesian
정확도	모든 컬럼	0.72	0.5	0.51	0.5	0.85	0.77
	Time/Amount 컬럼 제거	0.785	0.84	0.89	0.875	0.845	0.8
	V13,15,20 제거	0.795	0.825	0.89	0.905	0.845	0.795

5. 정리

- 총 6가지의 머신러닝 알고리즘을 30의 변수로부터 class를 0과 1로 분류하는 모델을 평가함
- Time 및 Amount 등의 불필요한 컬럼을 제거했을 때 예측 성능이 개선되는 것을 확인
- 또한 class별 histogram에서 나타난 것처럼 class와 상관없이 비슷한 패턴을 보이는 것은 feature로 활용되기 어렵다고 할 수 있음
- 즉, V13, V15, V20과 같이 class와 관련없는 컬럼을 추가로 제거 했을 때, 예측 정확도가 전반적으로 동등/향상 확인 → 적은 변수로 동등 혹은 향상 결과 도출
- training set에 따라, 알고리즘별 parameter들에 따라 성능이 달라질 수 있어 최적화된 결과는 아님
- 랜덤하게 샘플링을 하여 계산한 결과라 다른 샘플을 통해 분석하면 성능은 바뀔 수가 있어 충분한 반복이 필요함