[6주차 과제2] DA의 하이퍼파라미터 변경에 따른 성능결과 분석

당뇨병 발병 예측을 통한 분석

18011480 이진형 I 컴퓨터공학과

문제 분석

문제에서 주어진 데이터 셋(train.csv, test.csv)는 당뇨병에 관련된 개인의 상태를 모사한 데이터이다. 데이터 셋은 Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age의 8개 컬럼으로 구성되어 있고, 해당 데이터를 이용해 당뇨병 보유 여부를 예측하는 문제이다. 당 데이터를 이용, 이차 판별 분석(Quadratic Discriminant Analysis)를 통해 학습을 진행했다.

실험 조건

scikit-learn 문서에 따르면 Quadratic Discriminant Analysis의 파라미터는 priors, reg_param, store_covariance, tol 4개 가 존재한다.

실험

reg_param의 변화를 통해 진행한 실험의 조건과 결과는 다음과 같다.

reg_param	**Score**
0.0(default)	0.76623
0.1	0.75324
0.2	0.75324
0.3	0.74458
0.4	0.72294
0.5	0.70562
0.6	0.68831
0.7	0.69696
0.8	0.68398
0.9	0.67532
1.0	0.63203
baseline	0.59248

데이터 전처리를 위해 두가지 scaler를 통해 실험을 진행했다. (나머지 파라미터는 기본값)

Scaler	**Score**
StandardScaler	0.76623
RobustScaler	0.75324

실험 결과

위와 같은 조건으로 실험을 진행한 결과, 최고 점수인 0.76623을 기록한 파라미터는 다음과 같았다.
`reg_param = 0.0`

reg_param의 변화에 있어 최고 점수를 기록한 0.0은 scikit-learn에서 기본값으로 갖는 파라미터이다. 또한 0.1~1.0 의 변화를 갖고 실험한 결과 이번 문제에선 정규화 정도를 낮추는 것이 더 높은 점수를 기록한 것을 보였다.

개선사항

QDA를 통한 모델 개선은 기존 다른 학습 모델들과 대비해 수정할 수 있는 하이퍼파라미터의 개수가 적어 베이스라인 이상의 점수를 기록하는 등의 만족스러운 결과를 도출해 낼 수 없었다. 이 역시 데이터 전처리 과정을 통해 주어진 데이터 셋에서 주어진 독립 변수 들 중 변수들을 판별 변수로 선택해 모델 학습을 진행 할 것인지에 대한 공부가 필요할 것으로 보인다.