[6주차 과제1] DA의 하이퍼파라미터 변경에 따른 성능결과 분석

원자력발전소 상태 예측을 통한 분석

18011480 이진형 I 컴퓨터공학과

문제 분석

문제에서 주어진 데이터 셋(train.csv, test.csv)는 256차원의 발전소 모사 데이터이다. 데이터 셋은 발전소의 상태(A)를 나타내며, 데이터를 기반으로 변화한 후의 상태(B)를 판단하는 것이 목적이다. 당 데이터를 이용, 선형 판별 분석(Linear Discriminant Analysis)를 통해 학습을 진행했다..

실험 조건

scikit-learn 문서에 따르면 Linear Discriminant Analysis의 파라미터는 solver, shrinkage, priors, n_components, store_covariance, tol, covariance_estimator가 존재한다. 이번 실험에선 이들 중 solver, shrinkage, n_components 세 가지 파라미터의 변화를 통해 실험을 진행했다.
진행할 때 유의할 점으로 shrinkage 파라미터는 lsqr, eigen 이 두가지 solver 에서만 지원하고, n_components의 최대 값은 min(n_classes - 1 : 255, n_features : 197)인 197이다. 또한 n_components 파라미터는 데이터의 차원을 줄이는 것으로 transform을 지원하지 않는 lsqr solver에는 사용하지 않았다.

실험

진행한 실험의 조건과 결과는 다음과 같다.

Solver	Shrinkage	n_components	Score
svd	None	197	0.59248
svd	None	100	0.59248
svd	None	10	0.59248
lsqr	None	-	0.59226
lsqr	auto	-	0.58074
eigen	None	197	0.59226
eigen	None	100	0.59226
eigen	None	10	0.59226
eigen	auto	197	0.58074
eigen	auto	100	0.58074
eigen	auto	10	0.58074
QDA(with base parameters)			0.44772
baseline			0.59248

실험 결과

위와 같은 조건으로 실험을 진행한 결과, 최고 점수인 0.59248을 기록한 파라미터는 다음과 같았다.
`svd/None/197`
`svd/None/100`
`svd/None/10`
실험을 진행하면서 최대한 파라미터 변경에 따른 점수 차이를 확인할 수 있도록 각 파라미터 안에서 가장 크고 작은 값을 변수로 삼고자 했는데, 기대했던 수치 보다 점수 결과에 있어 차이가 미미해 아쉬웠다, 특히 최고 점수를 기록한 svd, None 옵션은 각각 scikit-learn에서 기본으로 제공하는 파라미터임을 보였다.

개선사항

LDA의 파라미터 변경으로 진행한 실험과, 추가적으로 QDA를 통해 진행한 실험 결과 모두 베이스라인보다 낮거나, 큰 점수 차이가 나지 않는 모습을 보여 보다 높은 점수를 위해서는 데이터 전처리를 통한 작업이 필요할 것이라고 예측한다.