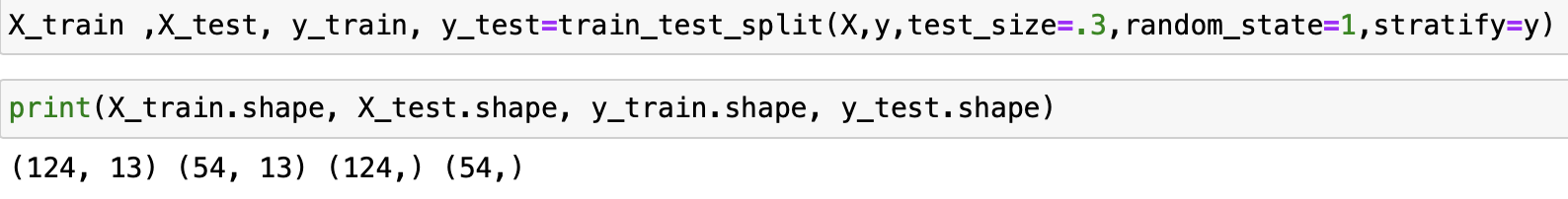
2019150445 신백록

1. 제5장의 wine data를 이용하여 linear, quadratic 그리고 naïve Bayesian을 적용하여 그 결과를 비교하고 제시하라.

Wine 데이터를 불러와 train set과 test set으로 split하였다. Test data가 54개, class가 3개로 data가 그리 적다고 느껴지지 않아 cross validation의 필요성을 느끼지는 못하였다.



그 후, LDA, QDA, Gaussian naïve bayes 모델에 데이터를 적합 시켜 train acc와 test acc를 확인하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

성능이 거의 비슷하지만 QDA가 LDA에 비해 overfitting이 되어있다는 것이 눈에 띈다. LDA는 linear임에 비해 QDA는 quadratic 모델이기에 추정해야하는 모수가 더 많아짐과 동시에 그에 따라 overfitting이 발생한 것으로 보인다. Wine data에 대해서는 QDA보다 성능도, 해석도 용이한 LDA를 사용하는 편이 좋아 보인다.

모델의 성능을 결정짓는 모델의 가정에 대해 확인을 해보았다. 일단 LDA는 가 정규분포를 따른다고 가정하고, class들의 covariance matrix가 동일하다고 가정한다. 정규성은 각각의 특성 변수에 대해 qq-plot을 그려 보았을 때, 만족한다고 보기 어렵다. Shapiro test를 해보아도 모든 클래스에서 univariate normal dist를 만족하는 특성 변수가 2-3개 밖에 존재하지 않는다는 것을 보았을 때 multivariate normal dist를 만족한다고 보기 어렵다.

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

등분산을 검정하기 위해 Class 별로 corr matrix를 구해보고, heatmap으로 나타내 보았다.

텍스트, 바닥, 체커, 검은색이(가) 표시된 사진

자동 생성된 설명 텍스트, 바닥, 체커, 바둑판식이(가) 표시된 사진

자동 생성된 설명 텍스트, 바닥, 실내, 검은색이(가) 표시된 사진

자동 생성된 설명

Matrix를 보면, class1의 correlation 색이 class3의 색보다 더 진하고 검정색 부분이 많은 것을 볼 수 있다. Class1은 음의 방향으로 더 많이 correlate 되어있는 반면, class3는 양의 방향으로 더 correlate 되어있는 것으로 보인다. 또한 각각의 변수들이 많이 correlate 되어있기에 Gaussian NB의 독립 가정도 만족하지 못하는 것으로 보인다. 그럼에도 모든 모델의 성능이 괜찮게 나온 것은, class 별로 well-separated 되어있어 그런 것으로 보인다.