통계적 머신러닝 6장 과제

1. 제5장의 wine data를 이용하여 linear, quadratic 그리고 naive Bayesian을 적용하여 그 결과를 비교하고 해석하라.

wine data는 3개의 class, {1,2,3}가 있고 특성변수는 13개가 있다. 데이터의 수는 178로 작다. 따라서 test 데이터의 accuracy는 test 데이터의 선택에 의존할 수 있어 유의해야한다.

(1) Linear discriminant analysis를 적용했을 때,
train data accuracy : 1.0
test data accuracy : 0.981
(2) Quadratic discriminant analysis를 적용했을 때,
train data accuracy : 1.0
test data accuracy : 0.963
(3) normal naive Bayes model을 적용했을 때,
train data accuracy : 1.0
test data accuracy : 0.981
특성변수에 count 데이터가 아닌 특성변수가 있기 때문에 naive Bayesian에서 multinomial을 사용하지 못한다.

linear, quadratic, naive Bayesian 모두 높은 정밀도를 가져 좋은 모형이다. (하지만 train 데이터 보다 test 데이터의 정밀도가 약간 높아 overfitting의 문제를 가지고 있다.) 따라서 특성변수에 대한 정규분포 가정이 타당한 것으로 보인다.
linear의 정밀도가 quadratic보다 높아 분산은 동질성이 있을 것이다. naive Bayesian의 정밀도도 높아 특성변수 서로 간에 독립(최소한 uncorrelated)을 만족한다.
linear 그리고 naive Bayesian가 quadratic 보다 정밀도가 높아 더 우수한 모형이다. linear와 naive Bayesian의 정밀도는 test 데이터를 기준, 0.981로 동일한 accuracy를 가지고 있다. 모형이 간단할수록 모형의 generalization 측면에서 우수하기 때문에 3개의 모형 중에서 **normal naive Bayes model**이 가장 우수하다.