\*이론

<주성분분석(Principal component analysis; PCA)>

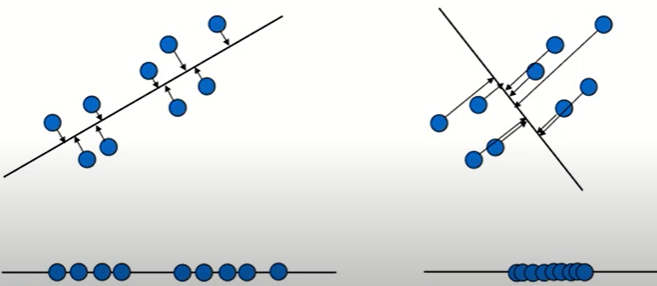
주성분분석은 고차원의 데이터를 저차원의 데이터로 환원시키는 방법을 말한다. 기존의 변수를 조합하여 서로 연관성이 없는 새로운 변수, 즉 주성분을 만들어 내는 분석 방법이다. 데이터를 한개의 축으로 사영시켰을 때 그 분산이 가장 커지는 축을 첫 번째 주성분, 두 번째로 커지는 축을 두 번째 주성분으로 놓이도록 데이터를 선형 변환한다. 사영했을 때 분산이 크다는 것은 원래 데이터의 분포를 잘 설명할 수 있다는 것과, 정보의 손실을 최소화할 수 있다는 것을 뜻한다. 이처럼 주성분 분석은 계산과 시각화가 용이하여 데이터를 쉽게 분석할 수 있다.

\*주성분 개수 선택의 기준

1.총 변이에 대한 공헌도 : culmulative proportion의 값이 85% 이상인 것까지 활용

2.개별 고유값의 크기 : 고유값이 1이상인 주성분 활용

3.screeplot을 통한 판단 : screeplot의 기울기가 완만해지는 부분 이전까지의 주성분 활용



출처: <https://www.youtube.com/watch?v=FhQm2Tc8Kic&list=PLpIPLT0Pf7IoTxTCi2MEQ94MZnHaxrP0j&index=11>

<Classificaiton>

LDA(Linear Discriminant Analysis)와 QDA(Quadratic Discriminant Analysis)는 주어진 데이터를 분류하는 데 사용되는 기법이다.

1) Linear Discriminant Analysis;LDA

클래스 간 분산 (Between class distance)을 최대화하거나 클래스 내 분산(within class distance)를 최소화하는 방식으로 데이터를 구분하는 데 필요한 축을 찾아내는 것을 목표로 한다. 이 방식은 각 클래스가 공통된 공분산을 가진다는 가정에서 실행된다. LDA는 입력 변수를 선형 조합으로 결합하여 데이터를 저차원 공간을 투영한다. 입력 변수의 차원을 줄이면서 클래스 간의 분리를 극대화한다. 이를 통해 클래스 간 구분이 더 쉬워지고 분류 정확도가 향상된다.

2) Quadratic Discriminant Analysis;QDA

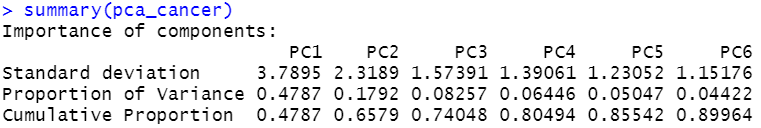
QDA는 클래스 간 분산과 클래스 내 분산이 공분산 행렬로 모델링된다. 각 클래스가 서로 다른 공분산을 가진다는 가정에서 수행된다. 이 방식은 각 클래스에서 공분산 행렬을 독립적으로 추정한다. 이러한 차이로 인해 2차원 이상의 비선형 결정 경계를 찾는다.

LDA와 QDA는 베이즈 정리를 기반으로 한다.

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명<코드 설명>

Screeplot을 그려보았을 때 다음과 같은 그래프를 확인할 수 있다. 기울기가 완만해지는 부분 이전까지의 주성분을 활용하고자 하였으나 정확히 한 지점을 판단할 수 없어 R에 내장된 prcomp을 이용하여 summary함수를 사용하였다. scale이 다른 문제로 인한 데이터의 왜곡을 피하기 위해 주성분 분석 시 scale=T을 사용하여 데이터를 표준화한다.

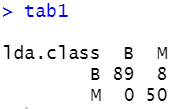


Cumulative Proportion 이 85% 이상이 되는 PC5까지 사용하기로 하자. 이는 원데이터의 약 89.96%보존한다고 이해할 수 있다. 5개의 주성분으로 차원 축소된 데이터를 생성하였다.

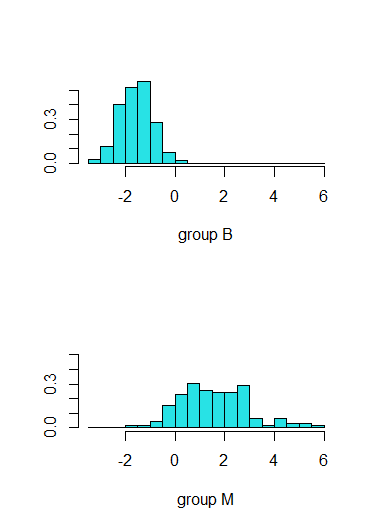
PCA분석을 통해 차원을 축소한 데이터로 Classification을 진행한다.

먼저, 다변량 정규분포를 따를 때, 공분산 행렬이 같다고 가정하는 LDA 방법을 사용해보자. 우선 Train data set과 Test data set으로 나눠야 하므로, 행개수의 70%를 비복원 랜덤추출하여 train 데이터로 사용하였고, 나머지는 test 데이터로 사용하였다. 다음 그림의 표를 통해 Test data set을 모형에 적합한 후 정분류율과 오분류율을 확인할 수 있다. 정분류율은 (89+50)/147=0.9455782, 즉 약 94.56%이고,

오분류율은 1-정분류율 =0.0544218, 즉 약 5.44%임을 확인할 수 있다.

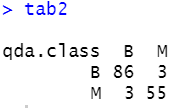


따라서 정분류율이 LDA모형 94.56%로 상당히 높다는 것을 알 수 있다.



<LDA 분류>

다음으로는 다변량 정규분포를 따를 때, 공분산 행렬이 다르다고 가정하는 QDA 방법을 사용해보자.



LDA 분류 방식과 동일한 방법으로 Train data set과 Test data set으로 나누어 분석하면 정분류율 약 95.92%, 오분류율 약 4.08%임을 확인할 수 있다.

이를 통해 PCA분석 후 데이터의 모형 적합 결과는QDA 방식이 LDA 방식에 비해 뛰어난 분류율을 가지고 있음을 알 수 있다.