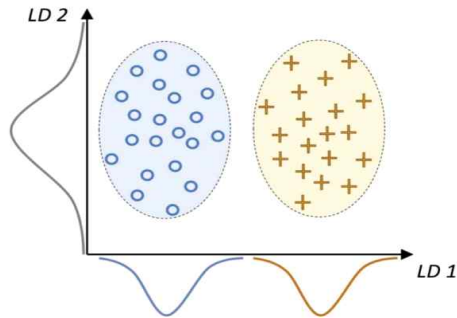


판별 분석(Discriminant Analysis)

: 2개 이상의 모집단(클래스)에서 가져온 표본들(샘플)이 지닌 정보(분포)를 이용해 이 표본들이 어느 모집단에 속하는지 결정할 수 있는 기준을 찾는 분석법



LD1: 투영된 선형 판별 벡터는 클래스를 잘 구분한다

LD2: 투영된 선형 판별 벡터는 클래스 판별 정보가 없어 좋지 않다

선형 판별 벡터 = LDA

1. 판별 변수(Discriminant variable)

: 샘플이 어느 집단에 속하는지 결정하기 위한 변수(판별력이 높은 독립 변수)
판별 변수 선택 조건

1. 판별 기여도

2. 서로 다른 독립 변수들과의 상관관계 → 상관관계가 적은 독립 변수 선택
ex) 상관관계가 높은 두 변수: 둘 다 X, 하나만 선택

2. 판별 함수(Discriminant function)

: 지도학습(supervised learning)이고 소속 집단을 알고 있는 변수들(X)의 판별력을 측정하고 새로운 대상을 어느 집단으로 분류할 것인지 예측한다

3. 판별 점수(Discriminant score)

: 소속을 판별하기 위해 대상의 판별 변수값을 판별 함수에 대입하여 구한 값

4. 표본의 크기

전체 표본(학습 데이터)의 크기 \geq 독립 변수의 개수 $\times 3$ (최소 $\times 2$)

종속 변수 Y의 개별 집단의 표본의 크기 중 최솟값 $>$ 독립 변수 X의 개수
(판별력 좌우하는 것은 표본의 개수가 아니라 집단의 표본 수 중 최솟값)

판별 분석의 단계

1. 판별 변수 찾기: 판별력 높은 독립변수

독립 변수 \sim Feature Engineering

2. 판별 함수 도출하기: 판별 변수들의 선형 결합

판별 함수 \sim 다양한 기계학습 방법론으로 학습하는 대상

: 판별 점수의 집단 간 변동과 집단 내 변동의 배율을 최대화

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Z : 판별 점수, β_0 : 판별 상수, X_1, X_2, \dots, X_p : 판별 변수, $\beta_1, \beta_2, \dots, \beta_p$: 판별 계수

3. 분류 정확도 분석하기: 학습 데이터 → 판별 함수

4. 클래스 예측하기: 테스트 데이터 → 판별 함수

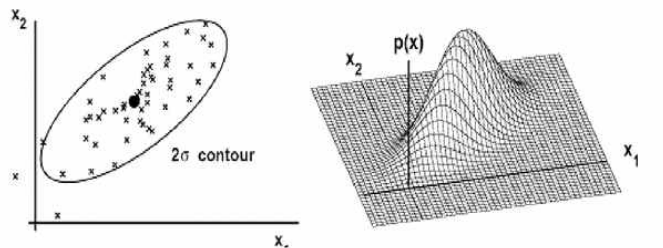
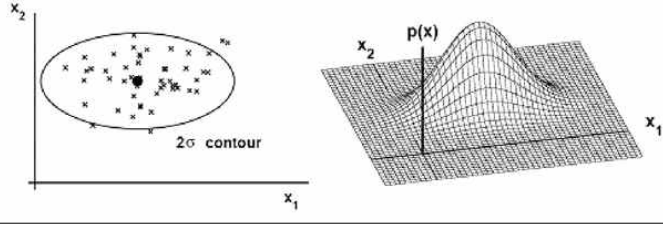
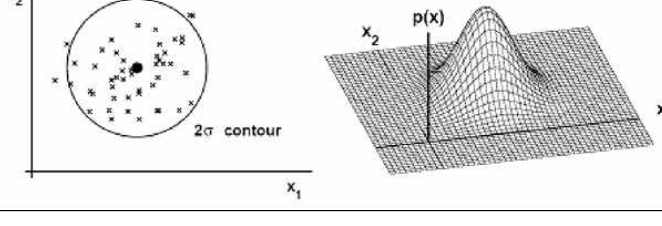
선형 판별 분석(LDA. Linear Discriminant Analysis)

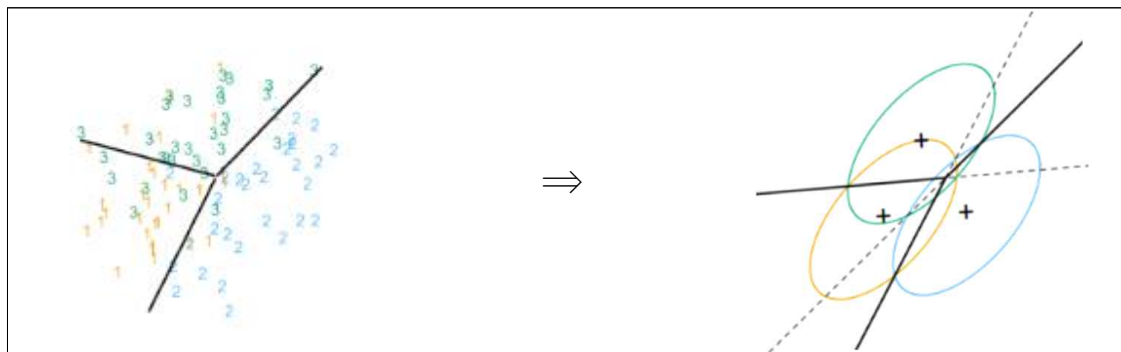
가정

1. 각 클래스 집단은 정규분포(normal distribution) 형태의 확률분포를 가진다

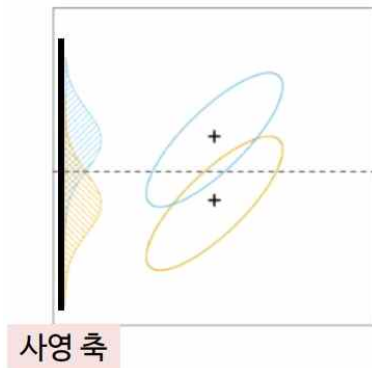
2. 각 클래스 집단은 비슷한 형태의 공분산(covariance)를 가진다

공분산: 2개의 확률 변수의 상관 정도

정규분포 = 가우시안 분포(Gaussian distribution)	
<p>완전 공분산 정규분포</p> $\Sigma = \begin{bmatrix} \sigma_1^2 & c_{12} \\ c_{12} & \sigma_2^2 \end{bmatrix}$	
<p>대각 공분산 정규분포</p> $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$	
<p>구형 공분산 정규분포</p> $\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$	



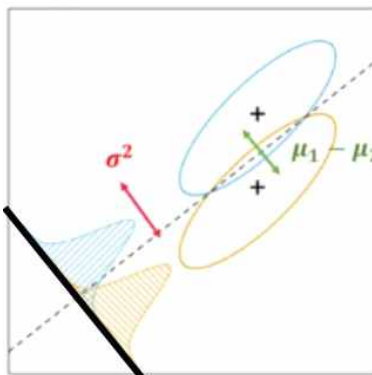
판별과 차원 축소의 기능



2차원(두 가지 독립변수)의 두 가지 변수를 갖는 데이터를 분류하는 문제에서 LDA는 먼저 1차원에 projection하여 차원을 축소한다

결정 경계

: 사영 축(실선)에 직교하는 축(점선)



결정 경계 선택 조건

1. 각 클래스 집단의 평균의 차이가 큰 지점
2. 각 클래스 집단의 분산이 작은 지점

\therefore 분산 대비 평균의 차이를 극대화하는 결정 경계
= 정사영 분포에서 겹치는 영역이 작은 결정 경계

장점

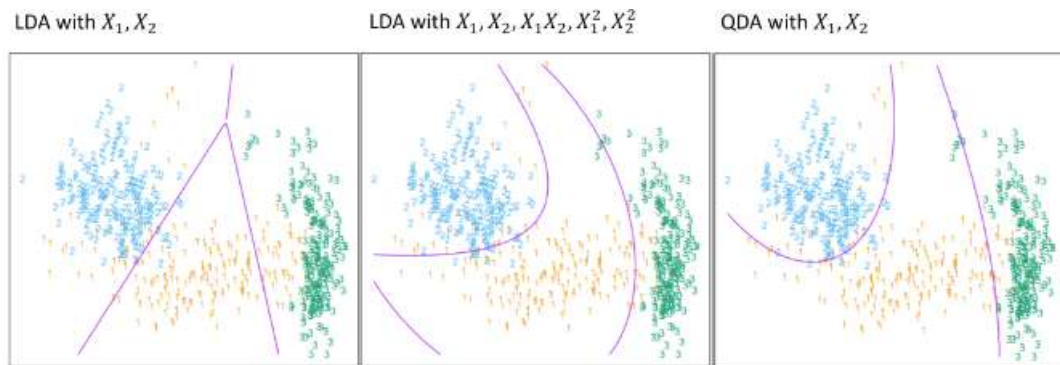
1. 변수 x 간 공분산 구조를 반영한다
2. 공분산 구조 가정이 살짝 위반되더라도 비교적 robust하게 동작한다

단점

1. 가장 작은 그룹의 샘플 수가 설명변수의 개수보다 많아야 한다
2. 정규분포 가정에 크게 벗어나면 잘 동작하지 못한다
3. 범주 y 사이에 공분산 구조가 많이 다른 경우(비선형)를 반영하지 못한다
→ 이차 판별 분석법(QDA)로 해결할 수 있다

이차 판별 분석(QDA, Quadric Discriminant Analysis)

범주의 수(k)와 관계없이 공통 공분산 구조에 대한 가정을 만족하지 못하면 적용
= Y의 범주 별로 서로 다른 공분산 구조를 가진 경우



가정: 클래스별 같은 공분산 가정: 클래스별 같은 공분산 가정: 클래스 별다른 공분산

LDA

선형 결정경계를 가정하여 서로 다른 공분산 분류는 어렵다

제곱을 한 추가적인 변수로 같은 공분산의 비선형 분류 가능하다

QDA ex) $y = \sigma_0 + \sigma_1 x_1 + \sigma_2 x_2 + \sigma_3 x_3$

장점; 서로 다른 공분산 데이터 분류 가능하다(비선형 분류 가능)

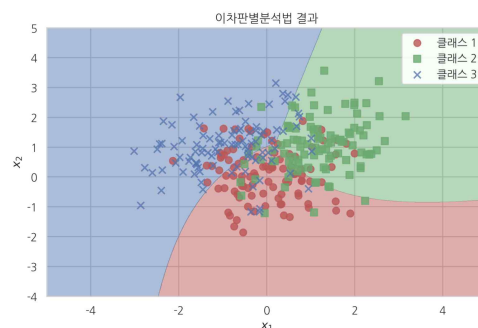
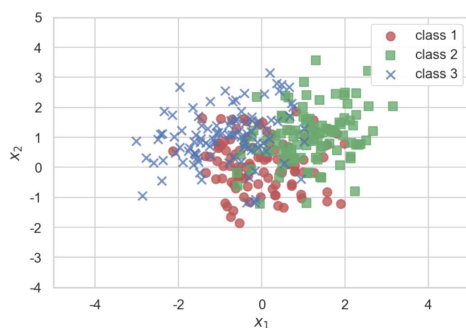
단점: 설명변수의 개수가 많으면 추정해야하는 모수가 많아진다(연산량 ↑)

ex) QDA: 클래스별 서로 다른 모수를 갖는 정규분포 분석

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mu_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.2 \end{bmatrix}$$

$$P(Y=1) = P(Y=2) = P(Y=3) = \frac{1}{3}$$



판별분석 api

LDA 클래스

[sklearn.discriminant_analysis.LinearDiscriminantAnalysis\(n_components=None, store_covariance=False\)](#)

n_components: 사영할 축의 차원(int, default = None)

store_covariance: 판별 함수에서 추정하고자 하는 모수를 보고자 하면
True(bool, default = False)

QDA 클래스

[sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis\(store_covariance=False\)](#)

store_covariance: 판별 함수에서 추정하고자 하는 모수를 보고자 하면
True(bool, default = False)