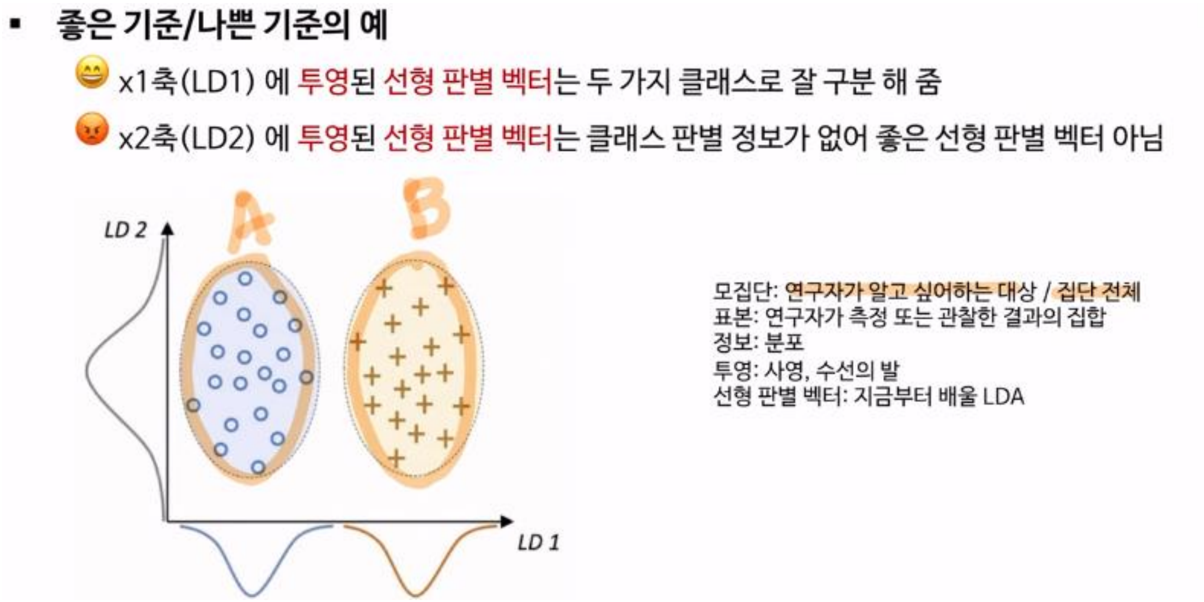
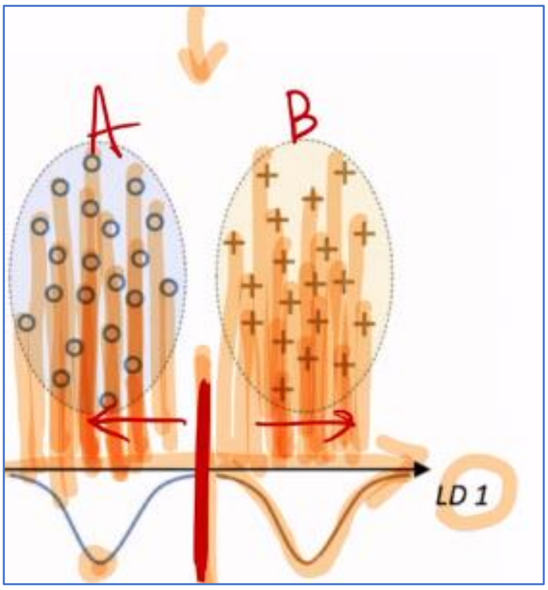
판별 분석 (Discriminant Analysis)

: 두 개 이상의 모 집단에서 추출된 **표본들이 지니고 있는 정보(데이터의 분포)를 이용하여** 이 표본들이 **어느 모집단에서 추출된 것인지를 결정할 기준을 찾는** 분석법

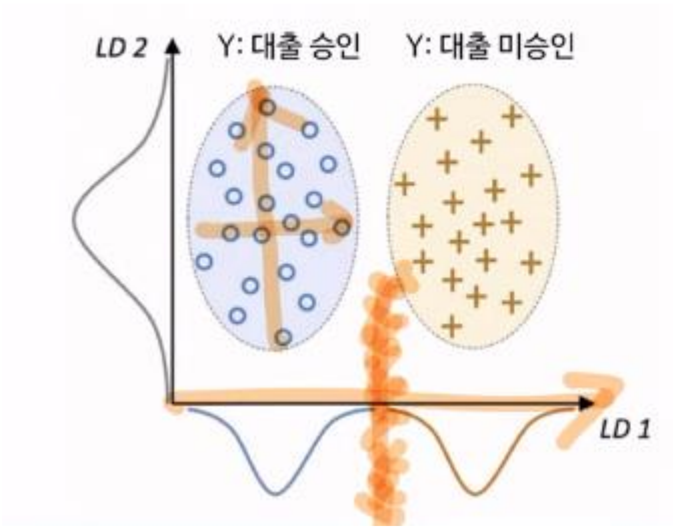




LD1으로 투영(Projection)했더니 클래스를 잘 구분해 줌

Ex) 은행에서 부동산 담보 대출을 행하고자 할 경우 채무자가 대출금을 갚을 것인가?

: 이 경우 과거 대출금을 반환하지 않은 사람의 정보 유형(연령, 소득, 결혼 유무 등)을 참고하여 담보 신청 시 신청자의 정보 유형을 과거의 유형과 비교하여 장래 변제 가능성을 파악 (학습 기반 분류 방법의 핵심)



X_1: 연령, X_2: 소득
Y : 대출 여부

판별 분석의 기초 개념

판별 변수(Discriminant Variable) : 어떤 집단에 속하는지 판별하기 위한 변수(X)로서 독립 변수 중 판별력이 높은 변수를 뜻함. 판별 변수를 선택하는데 판별 기여도 외에 고려해야 할 사항은 다름 독립 변수들 과의 상관관계이며, **상관관계가 높은 두 독립변수를 선택하는 것 보다는 두 독립변수 중 하나를 판별변수로 선택하고 그것과 상관관계가 적은 독립변수를 선택함으로써 효과적인 판별 함수를 만들 수 있음.**

판별 함수(Discriminant Function) : 판별 분석이 이용되기 위해서는 각 개체는 여러 집단 중에서 어느 집단에 속해 있는지(Y) 알려져 있어야 하며(supervised learning) 소속 집단이 이미 알려진 경우에 대하여 변수들(X)을 측정하고 이들 변수들을 이용하여 각 집단을 가장 잘 구분해 낼 수 있는 **판별식을 만들어 분별하는 과정**을 포함하게 됨.

즉, 판별 함수를 이용하여 **각 개체들이 소속 집단에 얼마나 잘 판별되는가에 대한 판별력을 측정하고 새로운 대상을 어느 집단으로 분류할 것이냐를 예측하는데 주요 목적**이 있음. (판별 함수를 통해서 LD1이라는 것을 찾겠다.)

판별 점수(Discriminant score) : 어떤 대상이 어떤 집단에 속하는지 판별하기 위하여 그 대상의 판별 변수들의 값을 판별 함수에 대입하여 구한 값.

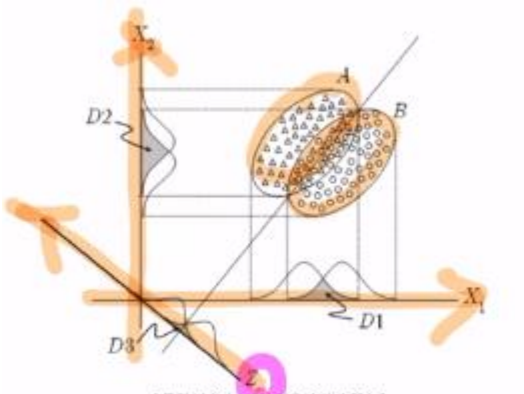
표본의 크기

- 전체 표본의 크기(학습 데이터의 수 N)는 통상적으로 독립변수의 개수(X, Dimension)보다 3배(최소 2배) 이상 되어야 함.

- 종속 변수의 집단 각각의 표본의 크기 중 최소 크기(가장 작은 class의 샘플 크기)가 독립 변수의 개수(x)보다 커야 함. (판별력을 좌우하는 것이 전체 표본의 수가 아니라 가장 적은 집단의 표본 수이기 때문)

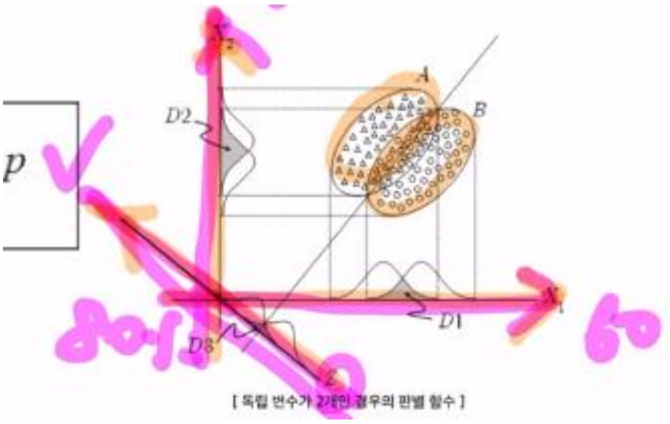
판별 분석의 단계 (판별 분석을 통한 분류기 설계 시작합니다)

1) 케이스가 속한 집단을 구분하는데 기여할 수 있는 **독립 변수 찾기**
2) 집단을 구분하는 기준이 되는 독립 변수들의 선형 결합 즉, **판별 함수 도출**하기

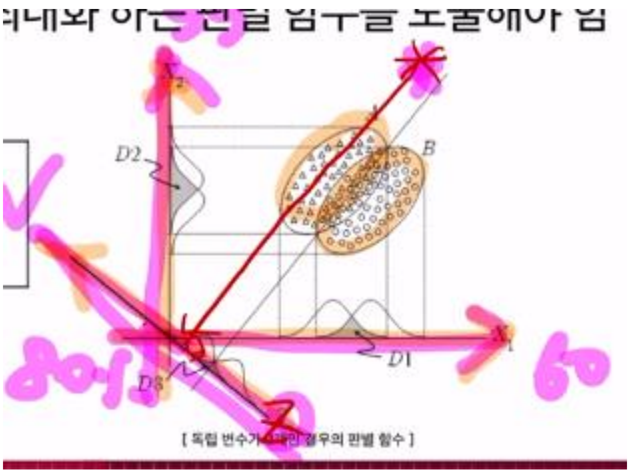


[독립 변수가 2개인 경우의 판별 함수]

3) 도출된 판별 함수에 의해 (학습 데이터) **분류의 정확도**를 분석하기



4) 판별 함수를 이용하여 새로운 케이스 (테스트 데이터)가 속하는 **클래스 예측**하기



독립 변수 <- Feature Engineering

판별 함수 <- 다양한 기계학습 방법론을 통해 학습하는 대상 (판별 점수의 집단간 변동과 집단 내 변동의 배율을 최대화하는 판별 함수를 도출해야 함. ???)

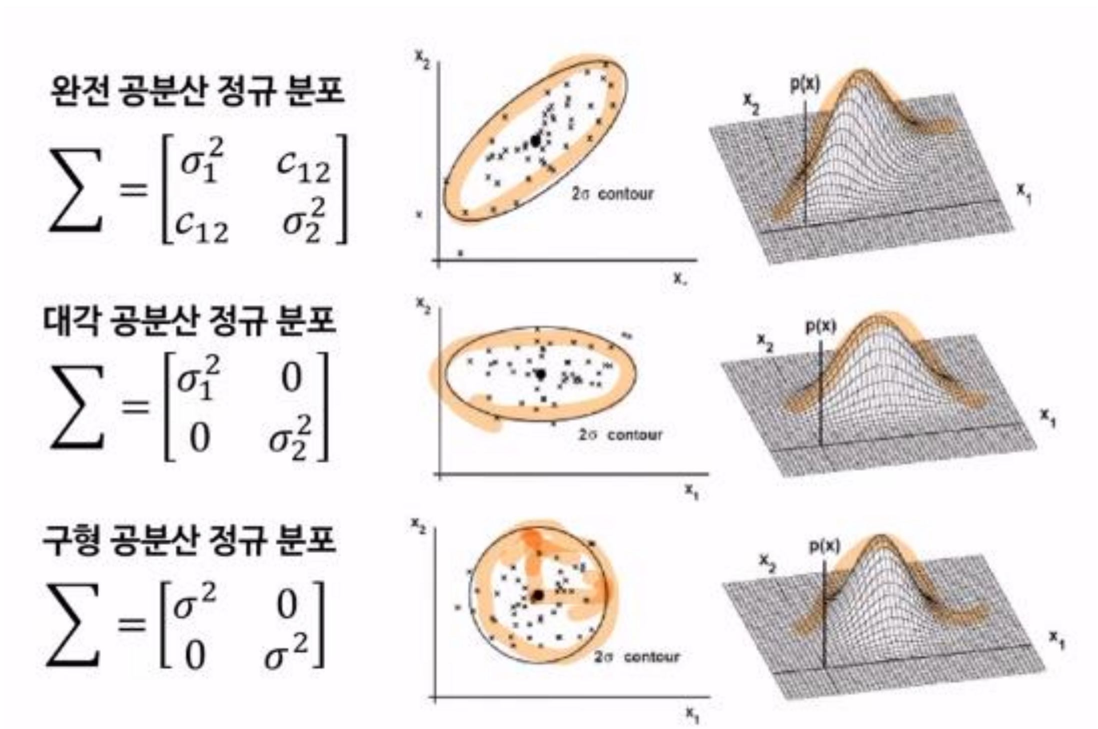
$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

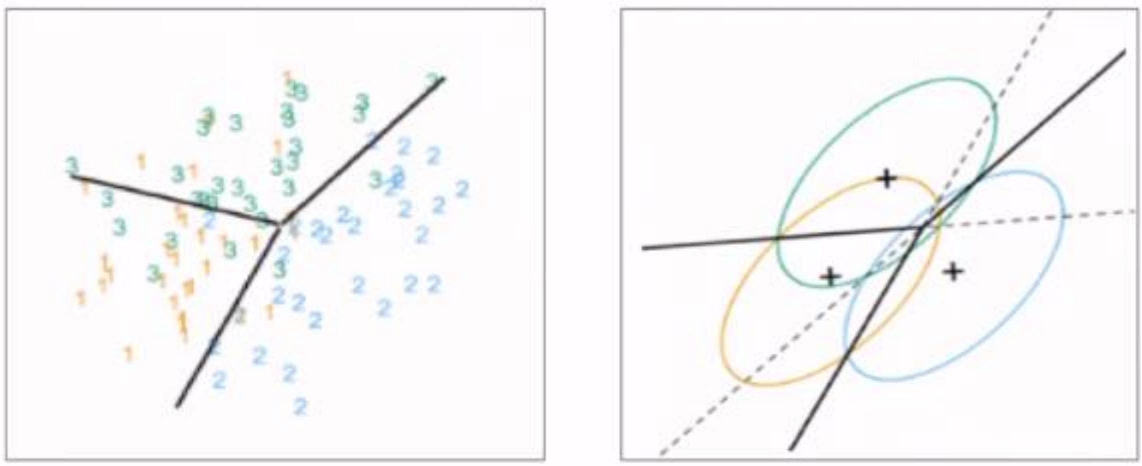
Z: 판별점수, β0: 판별상수, X1, X2, ..., Xp: 판별변수, β1, β2, ..., βp: 판별계수

선형 판별 분석 (Linear Discriminant Analysis) 배경

가정 (Assumptions) :

- 각 클래스 집단은 정규분포(Normal distribution) 형태의 확률 분포.
- 각 클래스 집단은 비슷한 형태의 공분산(Covariance) 구조를 가짐 (공분산 : 2개의 확률변수의 상관 정도를 나타내는 값)

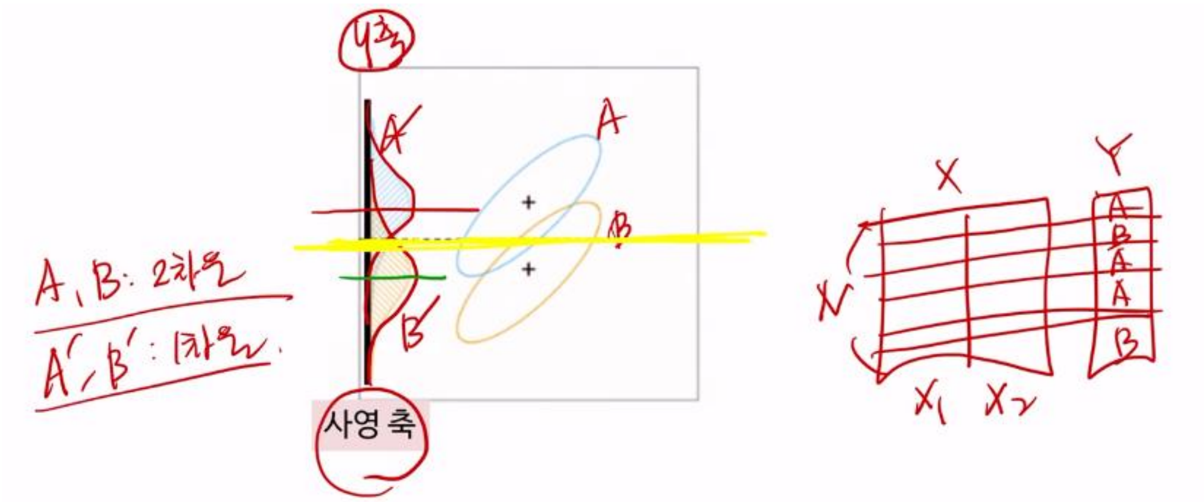




가정의 예시

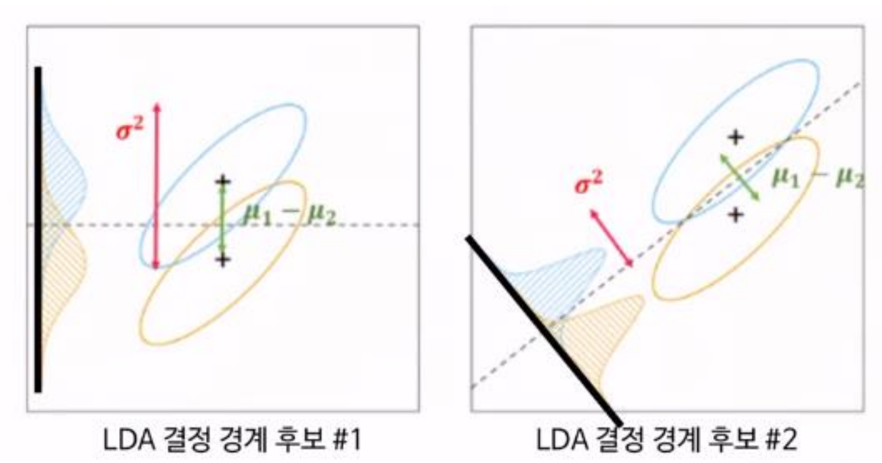
판별과 차원 축소의 기능 :

2차원 (두 가지 독립변수)의 두가지 범주(주황, 파랑)을 갖는 데이터를 분류하는 문제에서 LDA는 먼저 하나의 차원(1d)에 projection을 하여 차원을 축소시킴. (LDA는 차원 축소의 개념을 포함함, 2차원 자료들을 판별 축에 정사영 시킨 분포의 형태를 고려)



LDA의 결정 경계(Decision boundary)의 특징 :

- Projection 축(실선)에 직교하는 축(점선)
- 정사영(projection)은 두 분포의 특징이 아래의 목표를 달성해야 함
  - 각 클래스 집단의 평균의 차이가 큰 지점
  - 각 클래스 집단의 분산이 작은 지점
  - 즉, 분산 대비 평균의 차이를 극대화 하는 결정 경계를 찾고자 하는 것. (사영 데이터의 분포에서 겹치는 영역이 작은 결정 경계를 선택하면 됨)



위의 후보 둘 중 2번째 후보가 더 좋다!

선형 판별 분석 (LDA) 요약

- **결정 경계** : 분산 대비 평균의 차이를 극대화 하는 경계
- **가정**
  - ■ 각 클래스 집단은 정규분포(Normal distribution) 형태
  - ■ 각 클래스 집단은 비슷한 형태의 공분산(Covariance) 구조
- **장점**
  - ■ 변수(x) 간 공분산 구조를 반영
  - ■ 공분산 구조 가정이 살짝 위반되더라도 비교적 Robust 하게 동작
- **단점**
  - ■ 가장 작은 그룹의 샘플 수가 설명변수의 개수보다 많아야 함
  - ■ 정규분포 가정에 크게 벗어나는 경우 잘 동작하지 못함
  - ■ 범주 사이 (y)에 공분산 구조가 많이 다른 경우를 반영하지 못함.

    ➔ 이차 판별 분석 (QDA)를 통해 해결 가능

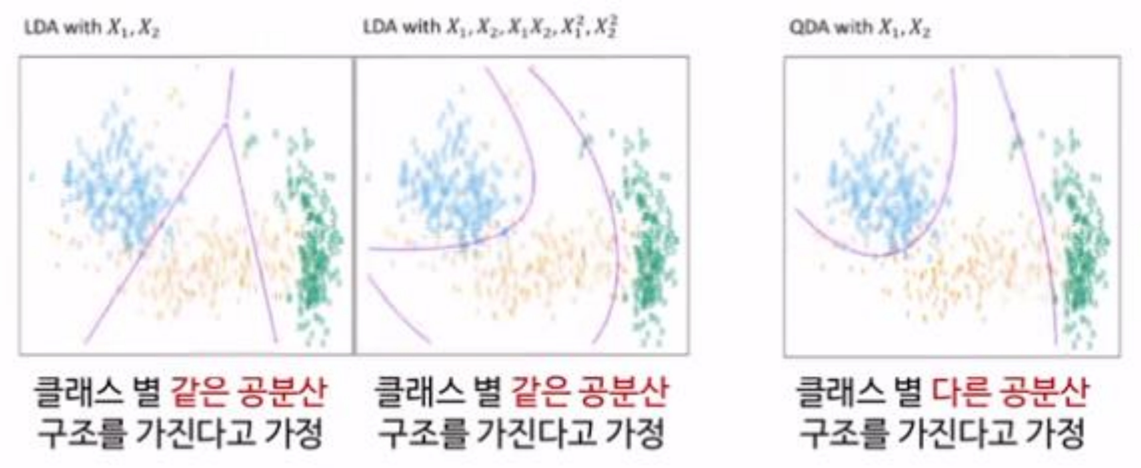
이차 판별 분석 (QDA)

이차 판별 분석의 정의 : K (범주의 수)와 관계없이 공통 공분산 구조에 대한 가정을 만족하지 못하면 QDA 적용. 즉 Y의 범주 별로 서로 다른 공분산 구조를 가진 경우에 활용 가능 (LDA의 두 번째 가정 불충족 시)

LDA와 QDA 비교 :

LDA의 결정 경계는 선형으로 가정하고 있어 서로 다른 공분산 분류에 어려움이 있음 (첫 번째 그림).

단, LDA도 같은 공분산의 비선형 분류 가능 (두 번째 그림, 변수의 제곱을 한 추가적인 변수들을 통해 보완, Data Augmentation)



QDA는 서로 다른 공분산 데이터 분류 가능 (세 번째 그림)

상대적 장점 : 비선형 분류 가능

QDA는 서로 다른 공분산 데이터 분류를 위해 샘플이 많이 필요.

상대적 단점 : 설명 변수의 개수가 많을 경우, 추정해야 하는 모수(Parameters)가 많아짐. 즉, 연산량이 큼

Ex) QDA : 클래스 별 서로 다른 모수를 갖는 정규분포 분석

- 예를 들어 y가 1,2,3 이라는 3개의 클래스를 가지고 각 클래스에서의 x의 확률분포가 다음과 같은 기대값 및 공분산 행렬을 가진다고 가정하자.

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

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

- y의 사전 확률은 다음과 같이 동일하다.

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