

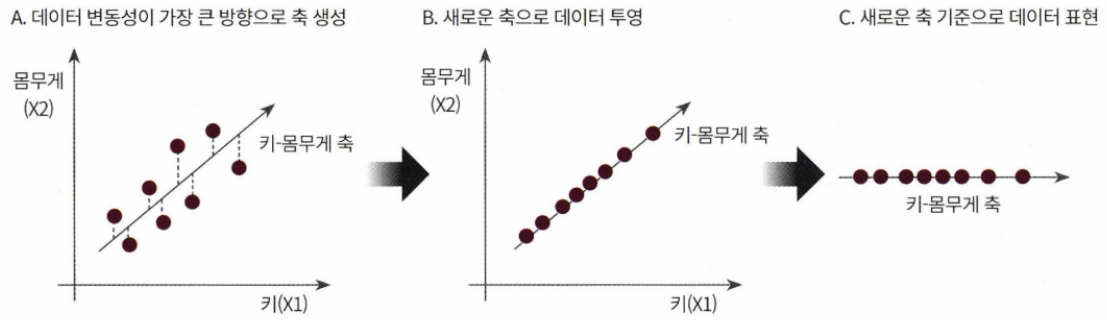
6장 차원 축소

01. 차원 축소(Dimension Reduction) 개요

- **차원 축소** : 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것
- → 일반적으로 차원이 증가할수록 희소한 구조를 가지게 된다. 또한 피처가 많을 경우 다중공선성 문제가 발생할 확률이 높다. 또한 이로 인한 모델의 예측 성능이 저하된다.
- **피처 선택(feature selection)**, **피처 추출(feature extraction)** : 차원 축소의 2가지 방법론
- **피처 선택** : 특정 피처에 종속성이 강한 불필요한 피처는 아예 제거하고, 데이터의 특징을 잘 나타내는 주요 피처만 선택하는 것
- **피처 추출** : 기존 피처를 단순 압축이 아닌, 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매핑해 추출하는 것 → 기존 피처가 전혀 인지하기 어려웠던 **잠재적인 요소(Latent Factor)**를 추출한다는데 그 의미가 있음
- 대표적인 차원 축소 알고리즘 : **PCA, LDA, SVD, NMF**
- ex) 이미지 : 이미지 변환과 압축을 통해 과적합 영향력을 작게 만들어 예측 성능을 끌어올릴 수 있다.
- 텍스트 : 단어들의 구성에서 숨겨 있는 시멘틱(Semantic) 의미나 토픽(Topic)을 잠재 요소로 간주하고 이를 찾아 낼 수 있다. (SVD, NMF등의 알고리즘이 사용된다.)

02. PCA(Principal Component Analysis)

- 가장 대표적인 차원 축소 기법이다.
- 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분(Principal Component)을 추출해 차원을 축소하는 기법이다.
- PCA는 가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원을 축소한다. (분산이 데이터의 특성을 가장 잘 나타낸다는 것으로 간주)



.

03. LDA(Linear Discriminant Analysis)

04. SVD(Singular Value Decomposition)

05. NMF(Non-Negative Matrix Factorization)