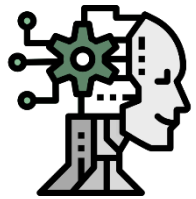
Chapter 6. Dimension Reduction



Kyungnam DA.
study group

1. 차원 축소 개요

`차원 축소` 란 ?

매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성

수백 개 이상의 피처로 구성된 데이터 세트의 경우 다중 공선성 등의 문제로 적은 차원의 모델보다 예측 신뢰도가 떨어질 수 있음.

피처를 차원 축소해 수를 줄이면 더 직관적으로 데이터를 해석할 수 있음

`차원 축소` 방법

Feature selection : 종속성이 강한 불필요한 피처를 제거하고, 데이터의 특징을 잘 나타내는 주요 피처만 선택

Feature extraction : 기존 피처를 저차원의 중요 피처로 압축해서 추출하는 것으로 새롭게 추출된 중요 특성은 기존의 피처와는 완전히 다른 것

1.1 차원 축소 알고리즘

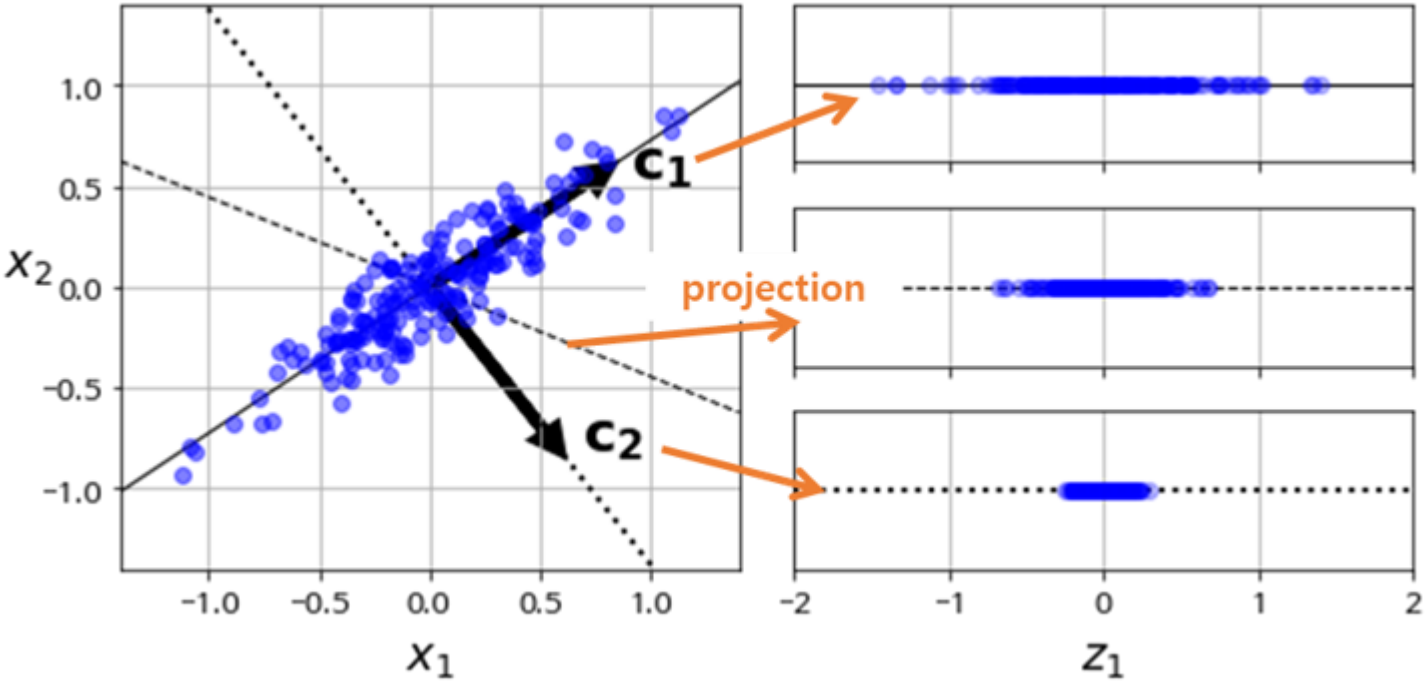
PCA (Principal Component Analysis)

여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분을 추출해 차원을 축소하는 기법

즉, 어떤 데이터들의 집합에서 가장 크게 해당 데이터를 구분 짓는 요소를 찾기 위한 분석 방법

데이터의 분산이 최대가 되는 초평면(Hyperplane)을 찾아 투영(projection) 시킴

입력 데이터의 공분산 행렬이 고유벡터, 고유값으로 분해가능하며, 고유백터를 통해 선형 변환한 방식



1) 데이터셋에서 분산이 최대인 축을 찾음
2) 첫번째 축과 직교하면서 분산이 최대인 두번째 축을 찾음
3) 첫번째 축과 두번째 축에 직교하고 분산을 최대한 보존하는 세번째 축을 찾음
4) 1~3과정의 방법으로 feature의 수만큼 축을 찾음

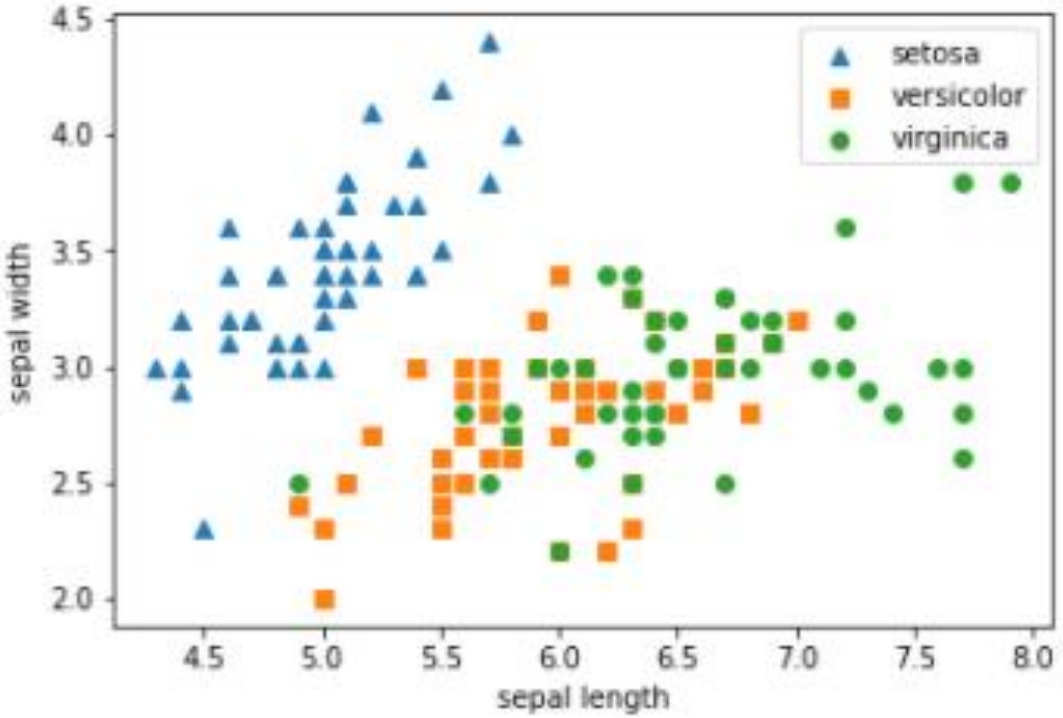
1.1 차원 축소 알고리즘

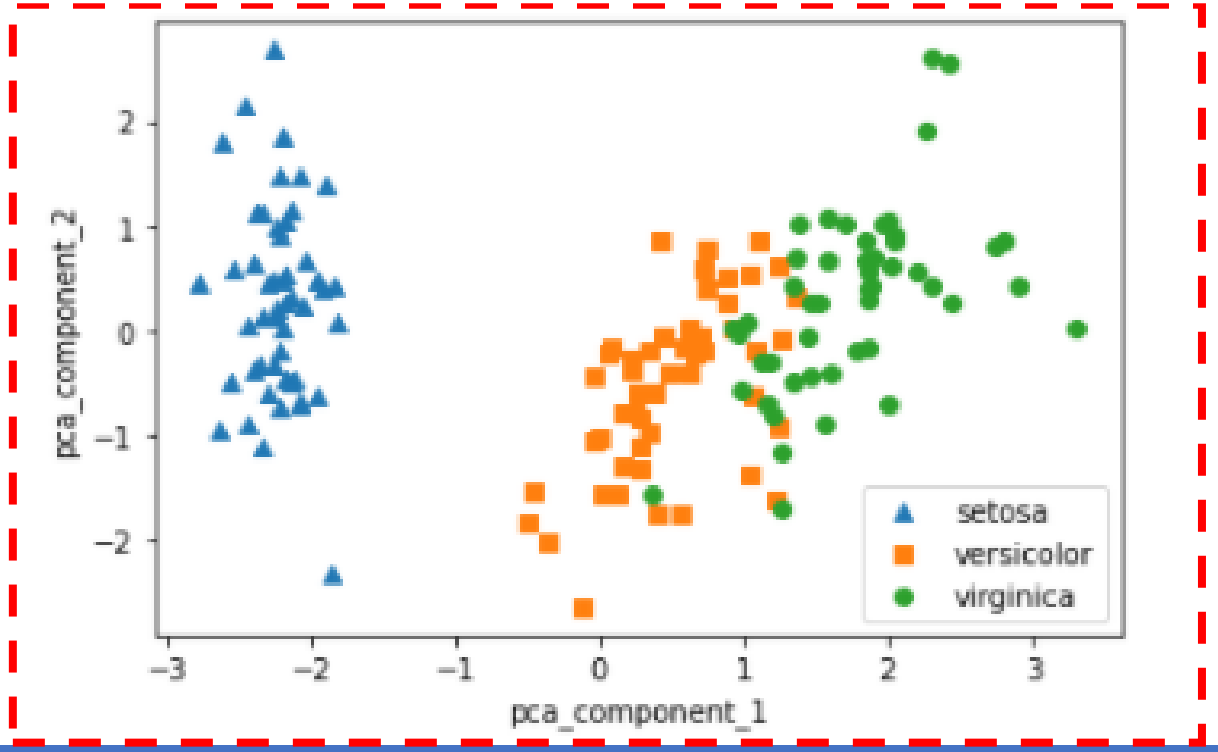
PCA (Principal Component Analysis)

`sepal width`, `sepal length` 조건만으로 분류를 알기에 시각적으로 충분하지 않음

따라서 PCA를 통해 여러 속성값을 축약하여 표현을 통해 분류를 파악

+ 여러 속성값을 연산해야 하므로 속성의 스케일에 영향을 받음 따라서 동일한 스케일로 변환이 필요함



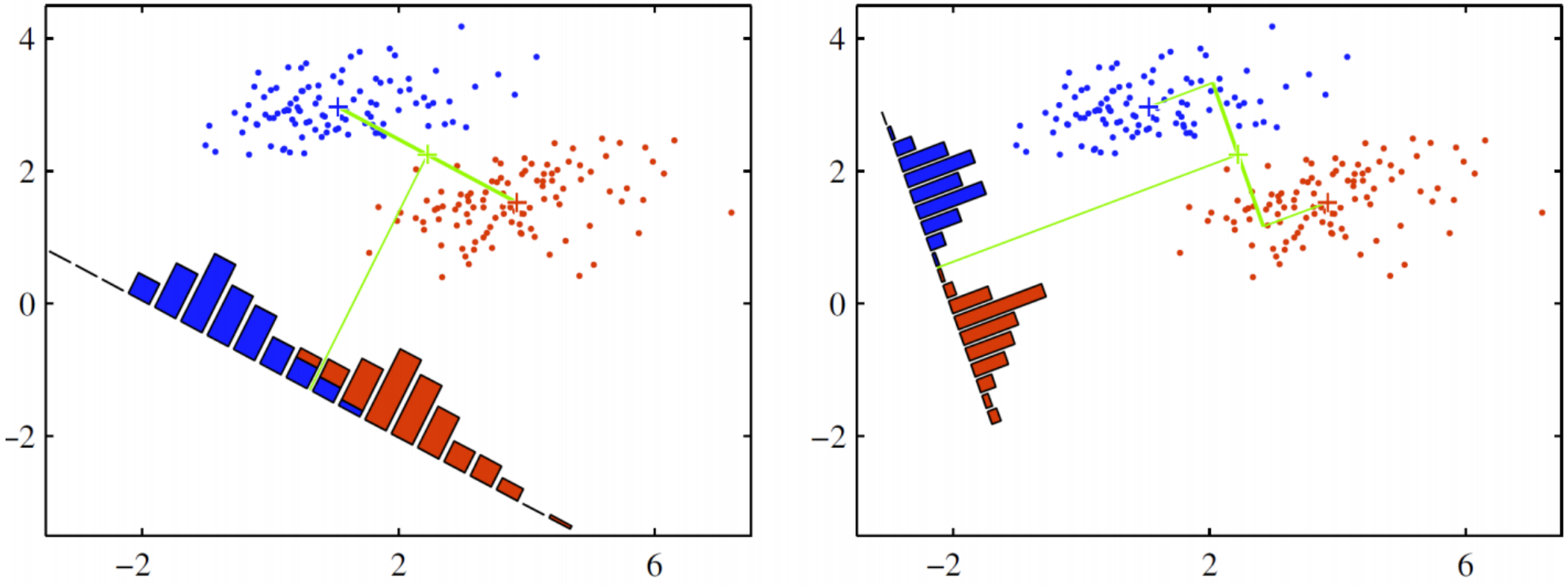


2. LDA (Linear Discriminant Analysis)

`선형 판별 분석` 이란 ?

PCA와 유사하게 분산을 최대화할 수 있는 축을 찾지만, 클래스를 분별할 수 있는 기준 축을 찾는 것

즉, 클래스 간 분산은 최대한 크게 가져가고, 클래스 내부의 분산을 최대한 적게 만드는 방식



3. SVD (Singular Value Decomposition)

`특이값 분해` 이란 ?

임의의 행렬 A를 정보량에 따라 여러 layer로 쪼개서 생각할 수 있게 해줌, 행과 열의 크기가 다른 행렬에서도 적용 가능 (고유값 분해는 정방 행렬에 대해서만 가능)

Singular Value Decomposition (SVD)

$$A = U \Sigma V^T$$

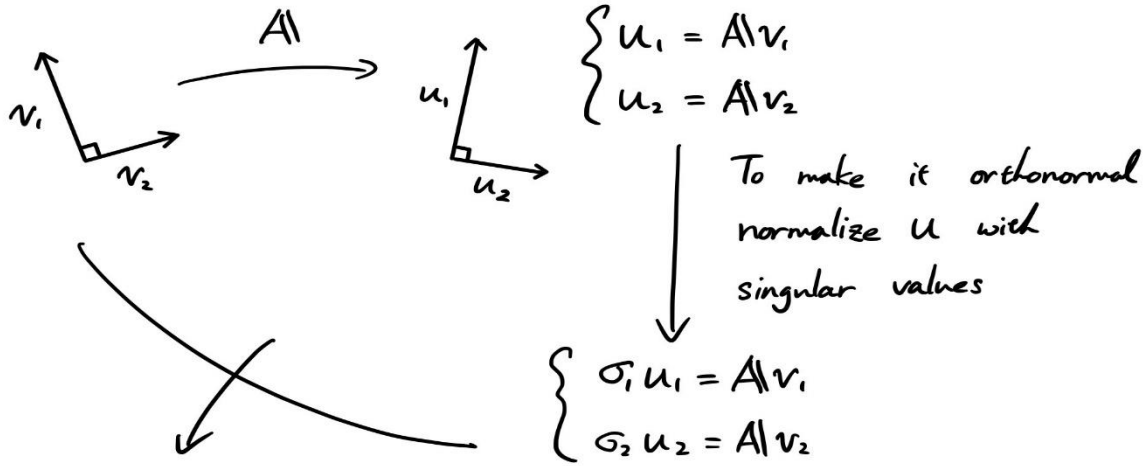
where $A \in \mathbb{R}^{m \times n}$ (Real or Complex matrix)

$U \in \mathbb{R}^{m \times m}$ (Real or Complex unitary matrix)

$\Sigma \in \mathbb{R}^{m \times n}$ (Rectangular diagonal matrix)

$V \in \mathbb{R}^{n \times n}$ (Real or Complex unitary matrix)

Intuitive explanation of SVD

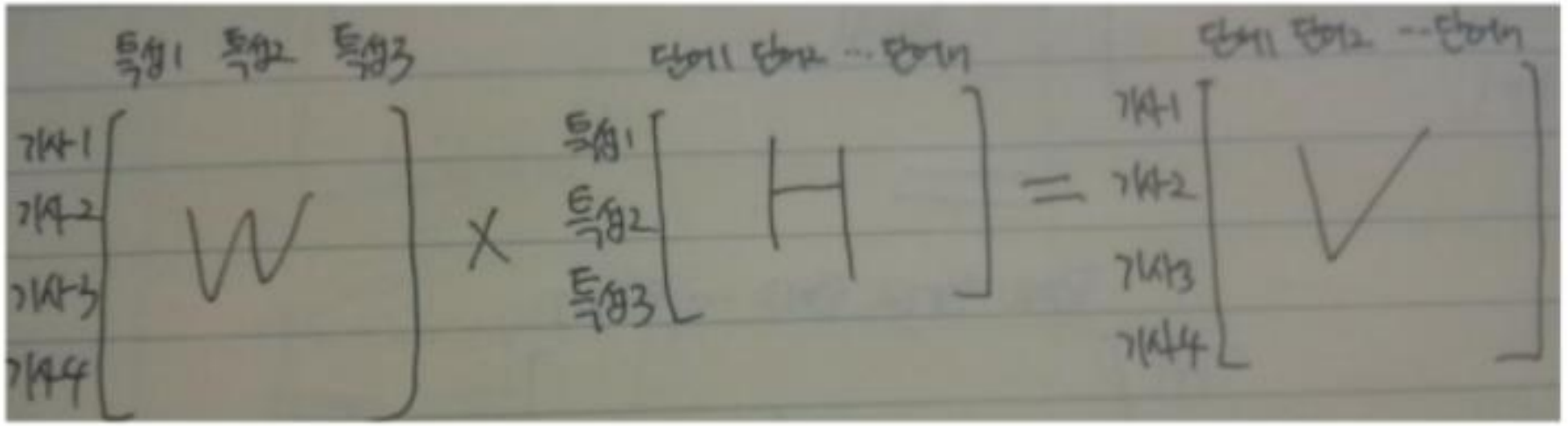


$$A[v_1 \; v_2] = [u_1 \; u_2]\begin{bmatrix}\sigma_1 & 0 \\ 0 & \sigma_2\end{bmatrix} \Rightarrow \boxed{AV = U\Sigma} \Rightarrow A = U\Sigma V^{-1} = U\Sigma V^T$$

4. NMF(Non-Negative Matrix Factorization)

`음수 미포함 행렬 분해` 이란 ?

음수를 포함하지 않은 행렬 V를 음수를 포함하지 않는 행렬 W와 H의 곱으로 분해하는 알고리즘



가중치 행렬:
- 각 특성이 기사에 얼마나 적합한가?
- 각 열에서 값이 높은 기사n개가 해당 특성을 갖는 기사
- 각 행에서 값이 높은 특성n개가 기사를 잘 반영하는 특성

특성 행렬:
- 각 단어가 특성에 중요한 정도
- 각 행에서 값이 높은 단어 상위n개가 해당 특성을 대표하는 단어가 됨 (즉, 주제가 됨)

* 특성의 개수 선택이 분류에 영향

Summary.

차원 축소는 피처의 개수를 줄이는 개념보다는 이를 통해 데이터를 잘 설명할 수 있는 잠재적인 요소를 추출하는데 큰 의미

PCA는 입력 데이터의 변동성이 가장 큰 축을 구하고, 다시 이 축에 직각인 축을 반복적으로 축소하려는 차원의 개수만큼 구한 뒤 축들에 투영해 차원을 축소하는 방법

LDA는 PCA와 매우 유사하지만 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축을 찾는 방식으로 차원 축소

SVD와 NMF는 매우 많은 피처 데이터를 가진 고차원 행렬을 두개의 저차원 행렬로 분리하는 행렬 분해 기법으로 행렬 분해를 수행하면 원본 행렬에서 잠재된 요소를 추출 가능