선형대수학팀

3팀
김진혁
김보근
노정아
심수현
이상혁

INDEX

1

주성분 분석

1 주성분 분석

차원의 저주

데이터 분석에서 **차원**이란 Feature의 개수를 의미

즉 차원의 증가는 설명변수가 늘어나는 것을 의미함

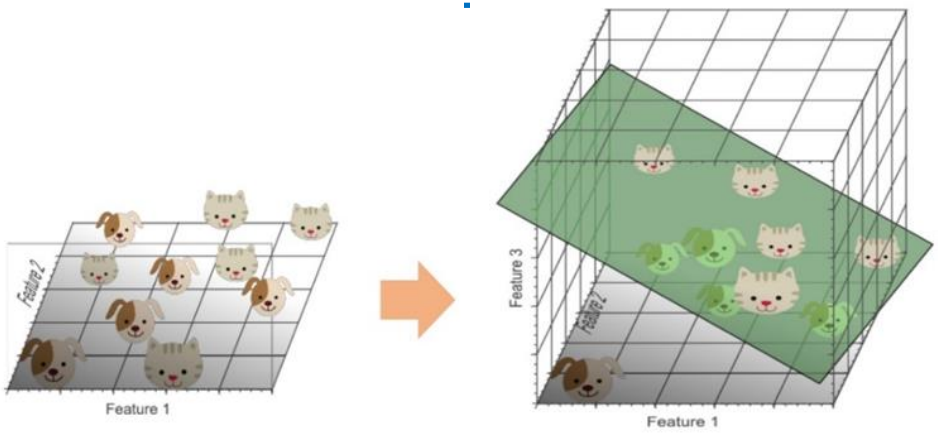
설명변수가 2개였을 때보다 차원의 증가로 분류가 쉬워짐

1 주성분 분석

차원의 저주

데이터 분석에서 **차원**이란 Feature의 개수를 의미

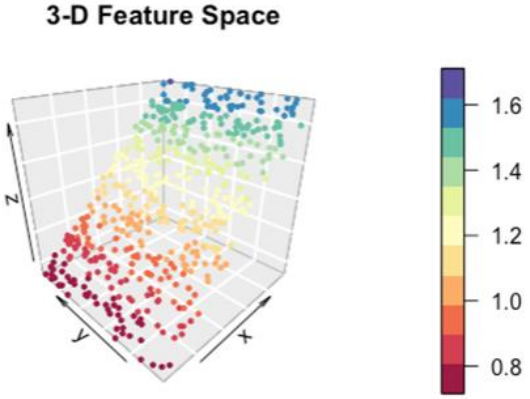
즉 차원의 증가는 설명변수가 늘어나는 것을 의미함



설명변수가 2개였을 때보다 차원의 증가로 분류가 쉬워짐

차원의 저주

변수가 많을수록 모델은 무조건 좋아질까?



데이터가 **3차원 공간에서 표현**되고 있지만, **한 평면** 위에 존재

2차원으로도 데이터를 다룰 수 있음

차원의 저주

차원의 저주

차원이 증가함에 따라 학습 알고리즘이 제대로 작동하지 않는 현상

계산량이 증가하여 데이터 **학습 속도**가 **떨어짐**

데이터 **밀도**가 **급격히 줄어듦**

↓

차원의 증가로 **공간의 낭비, 데이터 과적합** 등의 문제 발생

1 주성분 분석

차원의 저주

차원의 저주는 어떻게 해결하는가?

변수가 늘어나는 것이 왜 차원의 저주인가?

변수 선택

불필요한 변수를 제거하여

데이터 차원을 축소

3주차 회귀분석팀 클린업 참고

변수 추출

변수를 조합해 데이터를 잘 표현하는

중요 성분으로 **새로운 변수**를 추출

키, 몸무게, 머리길이, 눈 크기 등의 변수를 조합하여

가, 나, 다 라는 새로운 변수를 만드는 것

데이터가 3차원 공간에서 표현되고 있지만, 한 평면 위에 존재

1 주성분 분석

주성분 분석의 개념

주성분 분석(Principal Component Analysis)

변수 간의 **상관관계가 존재**하는 **다차원**의 데이터를

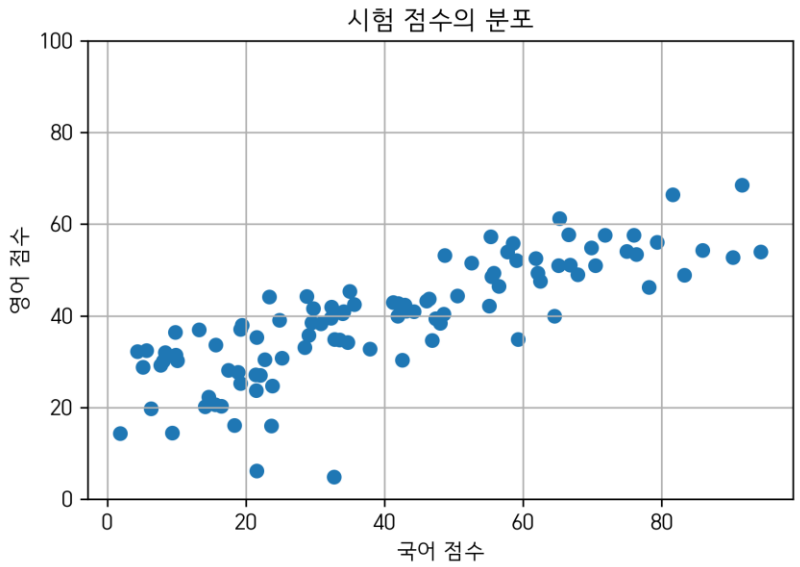
효율적으로 **저차원의 데이터로 요약**하는 차원축소 기법

데이터를 잘 설명해주는 주성분을 찾음

주성분이 이루는 공간으로 데이터를 정사영시켜 차원을 축소

1 주성분 분석

주성분 분석의 개념



100명의 국어성적을 x축에, 영어성적을 y축에 놓고 시각화

국어와 영어의 '종합 점수'를 만들고자 한다면 **어떤 기준**을 선택해야 하는가?

1 주성분 분석

주성분 분석의 개념

두 점수의 평균

국어 80점, 수학 60점이라면,

$$80 \times 0.5 + 60 \times 0.5$$

이를 벡터의 내적으로 표현하면,

$$\begin{pmatrix} 80 \\ 60 \end{pmatrix} \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

한 과목에 가중치

국어 80점, 수학 60점에
6:4 가중치를 적용하면,

$$80 \times 0.6 + 60 \times 0.4$$

이를 벡터의 내적으로 표현하면,

$$\begin{pmatrix} 80 \\ 60 \end{pmatrix} \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}$$

1 주성분 분석

주성분 분석의 개념

두 점수의 평균

국어 80점, 수학 60점이라면,

$$80 \times 0.5 + 60 \times 0.5$$

이를 벡터의 내적으로 표현하면,

$$\begin{pmatrix} 80 \\ 60 \end{pmatrix} \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

한 과목에 가중치

국어 80점, 수학 60점에

6:4 가중치를 적용하면,

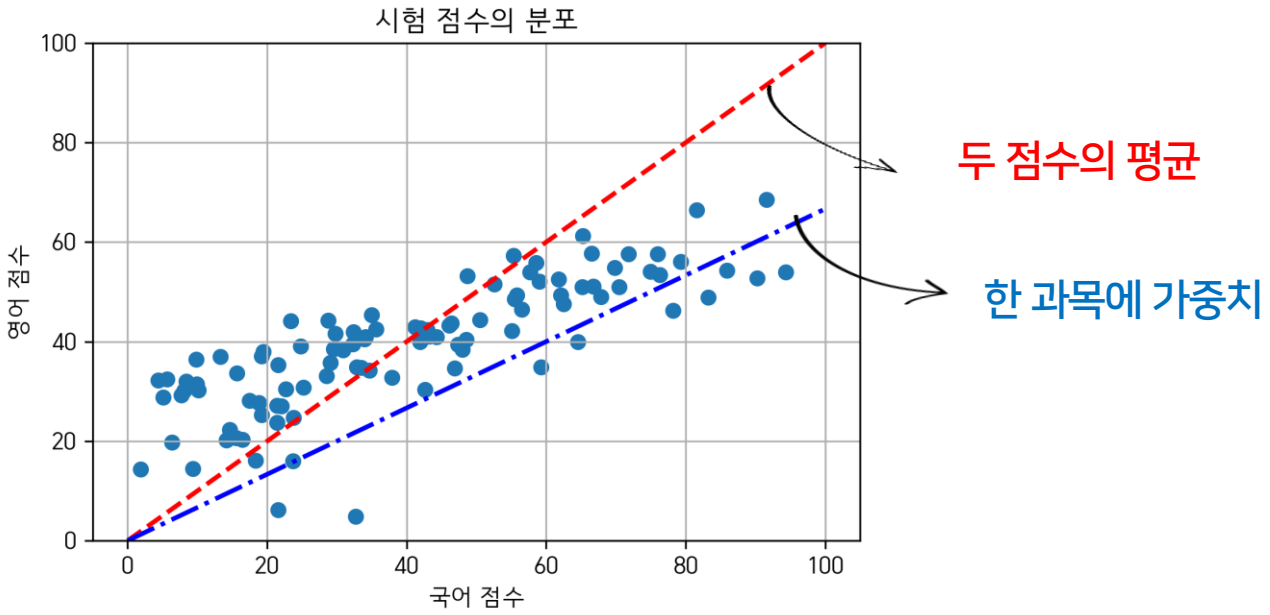
$$80 \times 0.6 + 60 \times 0.4$$

이를 벡터의 내적으로 표현하면,

$$\begin{pmatrix} 80 \\ 60 \end{pmatrix} \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}$$

1 주성분 분석

주성분 분석의 개념



종합 점수를 얻는 방식을 수학적으로는 벡터를 특정 비율을 표현하는

벡터에 내적(정사영)하는 문제로 환원

주성분 분석의 개념

어떤 벡터에 **내적(정사영)**하는 것이 최적의 결과를 낳는가?

내적의 대상이 되는 벡터를 찾을 때,

데이터 분포의 중심을 **중심축(Pivot)**으로 움직이는 벡터를 찾는 것이 좋지 않을까?

주성분 분석의 개념

이 문제에 대한 해결책은 **공분산 행렬**로부터 찾을 수 있음

공분산 행렬(Covariance Matrix)

공분산 행렬

변수들의 상관 정도인 **공분산**을 행렬로 나타낸 것

$$Cov(X,Y) = E\left[(X - \mu_x)(Y - \mu_y)^T\right] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mn} \end{pmatrix}$$

공분산 행렬(Covariance Matrix)

공분산 행렬의 수식적 의미

$$X = \begin{pmatrix} | & | & \cdots & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & \cdots & | \end{pmatrix} \in R^{n \times d}$$

위 행렬의 각 열의 평균이 0이라면,

데이터 분포의 중심을 중심축(pivot)으로 움직이는 벡터를 찾기 쉬움

각 열의 평균이 0이 아니라면,

평균을 빼서 각 열의 평균이 0인 행렬을 만듦

공분산 행렬(Covariance Matrix)

공분산 행렬의 수식적 의미

행렬 X를 이용하여 공분산 행렬 계산

$$X^T X = \begin{pmatrix} - & X_1 & - \\ - & X_2 & - \\ \vdots & \vdots & \vdots \\ - & X_d & - \end{pmatrix} \begin{pmatrix} | & | & \cdots & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & \cdots & | \end{pmatrix}$$

$$= \begin{pmatrix} dot(X_1, X_1) & dot(X_1, X_2) & \cdots & dot(X_1, X_d) \\ dot(X_2, X_1) & dot(X_2, X_2) & \cdots & dot(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ dot(X_d, X_1) & dot(X_d, X_2) & \cdots & dot(X_d, X_d) \end{pmatrix}$$

두 벡터의 내적을 의미

공분산 행렬(Covariance Matrix)

공분산 행렬의 수식적 의미

행렬 X를 이용하여 공분산 행렬 계산

$$X^T X = \begin{pmatrix} - & X_1 & - \\ - & X_2 & - \\ \vdots & \vdots & \vdots \\ - & X_d & - \end{pmatrix} \begin{pmatrix} | & | & \cdots & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & \cdots & | \end{pmatrix} =$$

$X^T X$ 행렬을 통해 i번째 변수와 j번째 변수가

얼마나 닮았는지 알 수 있음

$$\begin{pmatrix} dot(X_1, X_1) & dot(X_1, X_2) & \cdots & dot(X_1, X_d) \\ dot(X_2, X_1) & dot(X_2, X_2) & \cdots & dot(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ dot(X_d, X_1) & dot(X_d, X_2) & \cdots & dot(X_d, X_d) \end{pmatrix}$$

두 벡터의 내적을 의미

공분산 행렬(Covariance Matrix)

공분산 행렬의 수식적 의미

샘플이 커질수록 내적값이 계속 커지는 것을 방지하기 위해서 n으로 나눠줌

$$\frac{X^T X}{n} = \frac{1}{n}\begin{pmatrix} dot(X_1, X_1) & dot(X_1, X_2) & \cdots & dot(X_1, X_d) \\ dot(X_2, X_1) & dot(X_2, X_2) & \cdots & dot(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ dot(X_d, X_1) & dot(X_d, X_2) & \cdots & dot(X_d, X_d) \end{pmatrix}$$

비대각 요소를 보면 **대칭 행렬**임을 알 수 있음

따라서 대칭 행렬의 성질에 따라 **고유값 분해**가 가능

*항상 고유값 대각화, 직교행렬로 대각화 가능

공분산 행렬(Covariance Matrix)

공분산 행렬은 다음과 같이 간단히 표현 가능

$$\sum = \frac{1}{n} X^T X$$

공분산 행렬(Covariance Matrix)

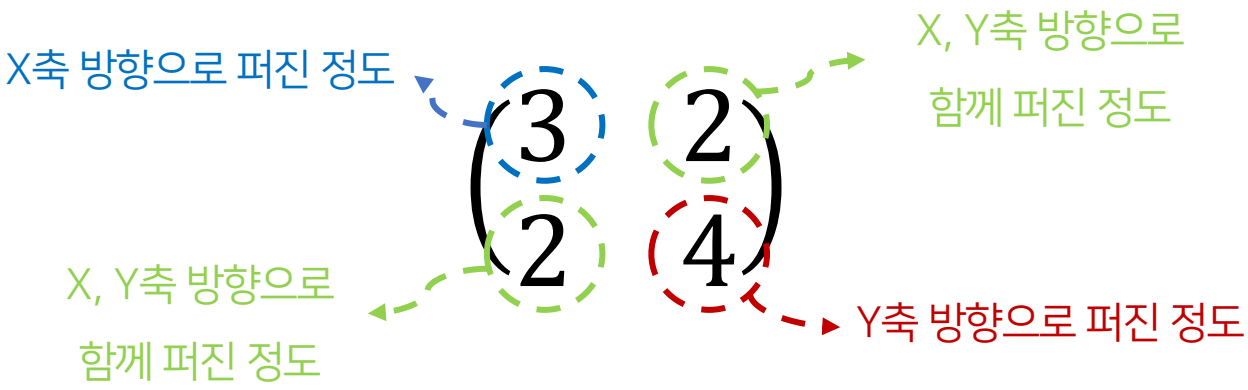
공분산 행렬의 기하학적 의미

공분산 행렬은 **데이터의 구조를 설명**하여 **변수 간의 관계를 파악**하는데 도움을 줌

↓

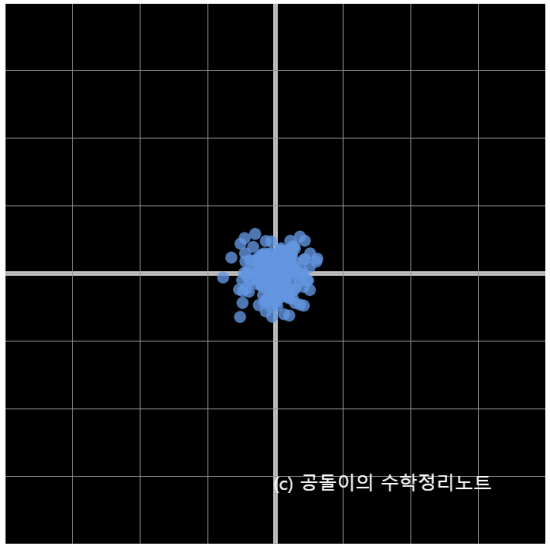
하나의 **선형변환**으로 생각

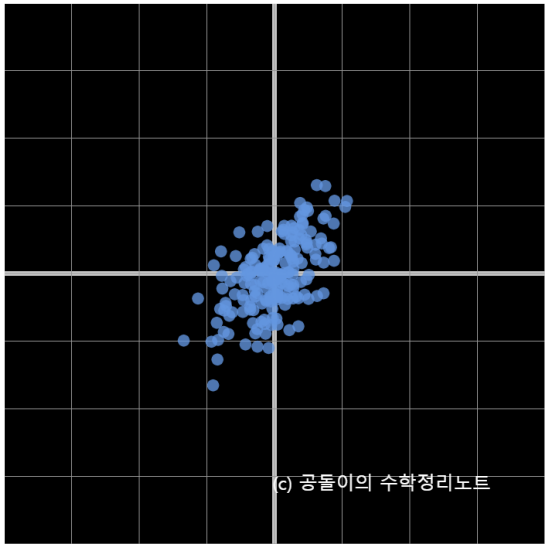
공분산 행렬을 통해 데이터를 변환하면 분산과 공분산만큼 공간이 변화함



공분산 행렬(Covariance Matrix)

공분산 행렬의 기하학적 의미





공분산 행렬 $\begin{pmatrix} 3 & 2 \\ 2 & 4 \end{pmatrix}$ 을 통해 데이터를 오른쪽으로 변환시킴

데이터가 어느 방향으로 어떻게 분포되어 있는지 알 수 있음

공분산 행렬(Covariance Matrix)

데이터의 분산과 정보량의 관계

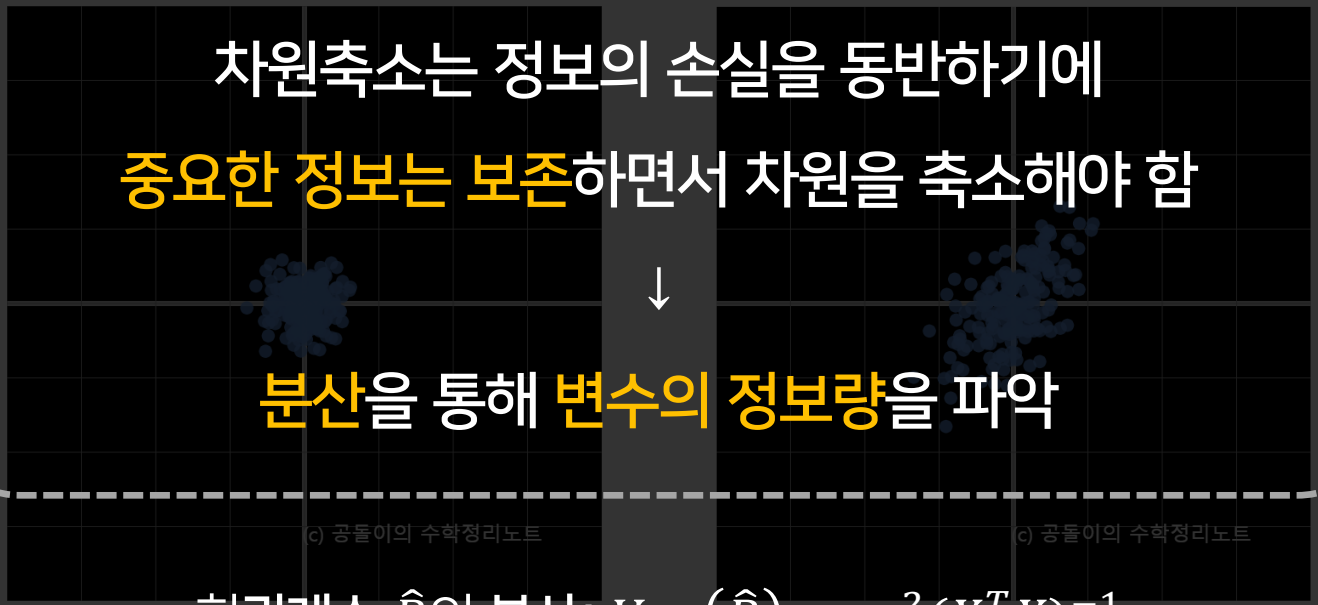
공분산 행렬의 기하학적 의미

차원축소는 정보의 손실을 동반하기에

중요한 정보는 보존하면서 차원을 축소해야 함

↓

분산을 통해 변수의 정보량을 파악

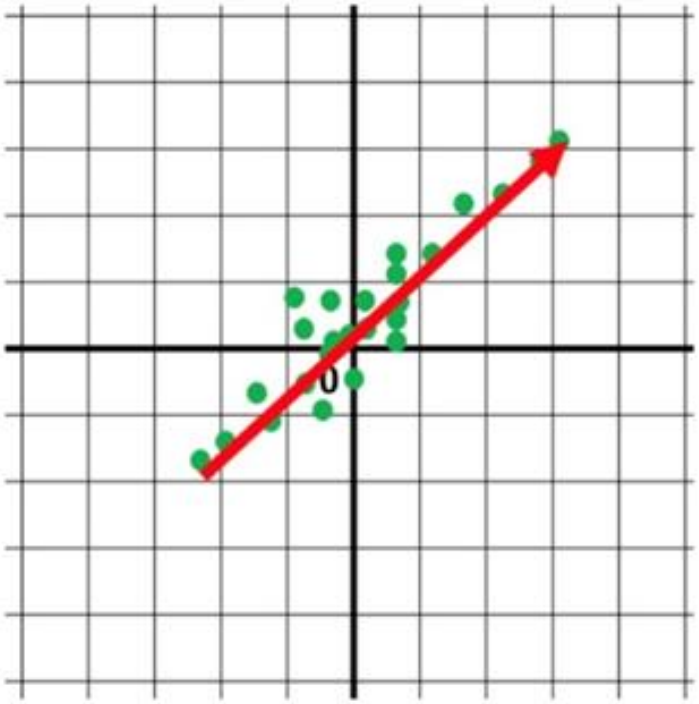


회귀계수 $\hat{B}$의 분산: $Var(\hat{B}) = \sigma^2(X^TX)^{-1}$

X의 분산이 크다면 X^TX 값이 크고 $Var(\hat{B})$이 작음

설명 변수의 분산이 크면 좋은 예측이 가능

주성분 분석



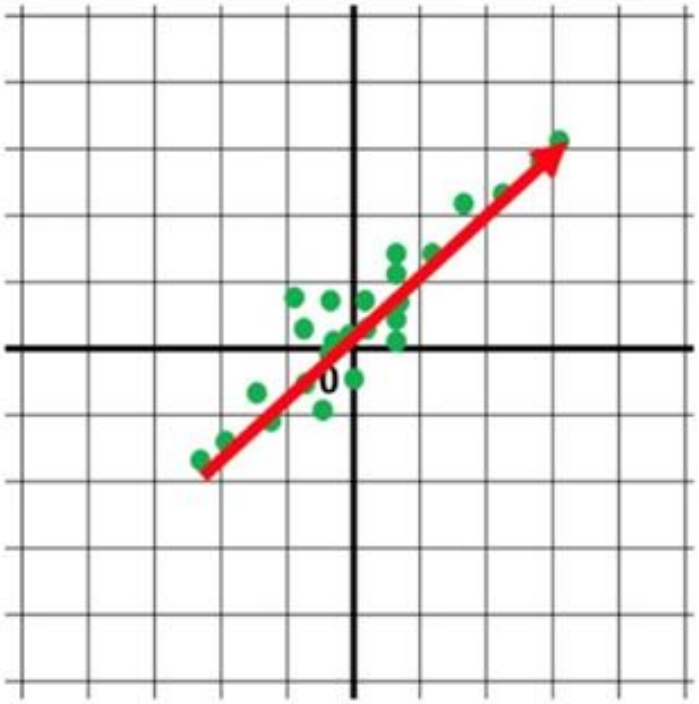
주성분(PC) 찾기

공분산 행렬 $\begin{pmatrix} 3 & 2 \\ 2 & 4 \end{pmatrix}$에서 빨간색 벡터 방향으로 projection을 내리는 것이 데이터의 분산을 제일 잘 보존

↓

이 빨간색 벡터는 어떻게 찾는가?

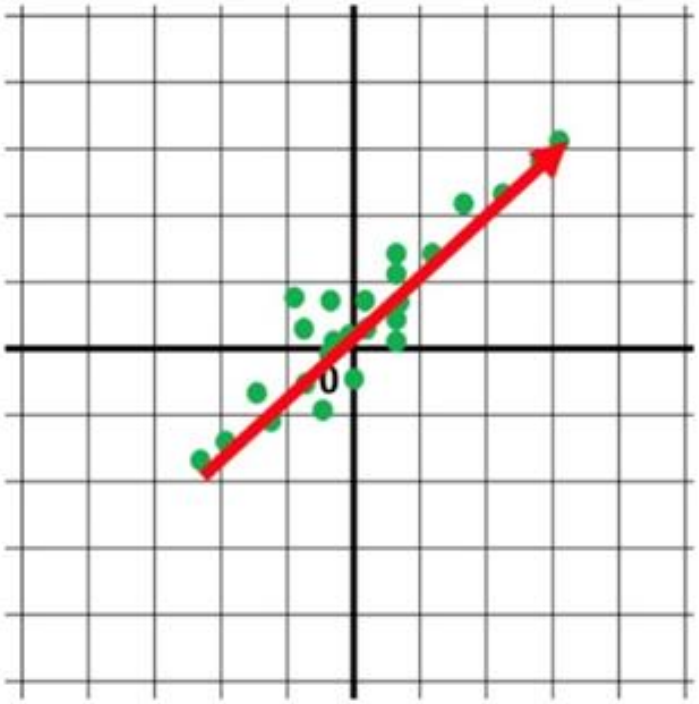
주성분 분석



대각선 방향으로 데이터를 잡고 늘리는
형식으로 선형 변환이 적용됨

빨간색 벡터의 방향은 바뀌지 않음

주성분 분석



대각선 방향으로 데이터를 잡고 늘리는
형식으로 선형 변환이 적용됨

빨간색 벡터의 방향은 바뀌지 않음

↓

고유값과 **고유벡터**를 이용

주성분 분석

고유벡터는 선형 변환 이후에도 크기만 변함

따라서 고유벡터는 선형 변환의 고정된 축

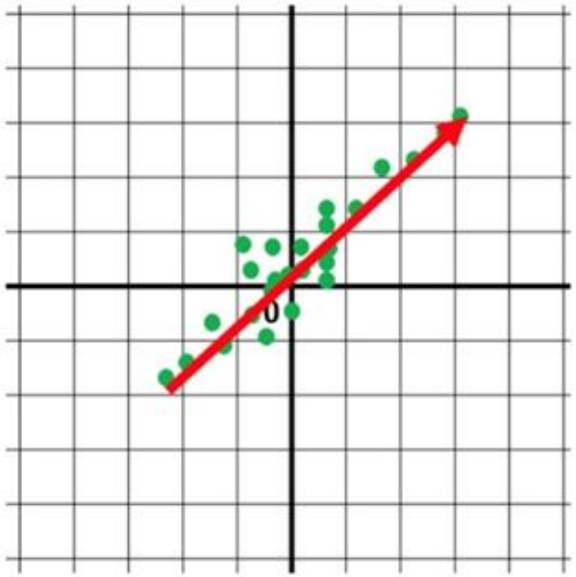
공분산 행렬의 고유벡터는 데이터가

어떤 방향으로 분산되어 있는지 나타냄

주성분 분석

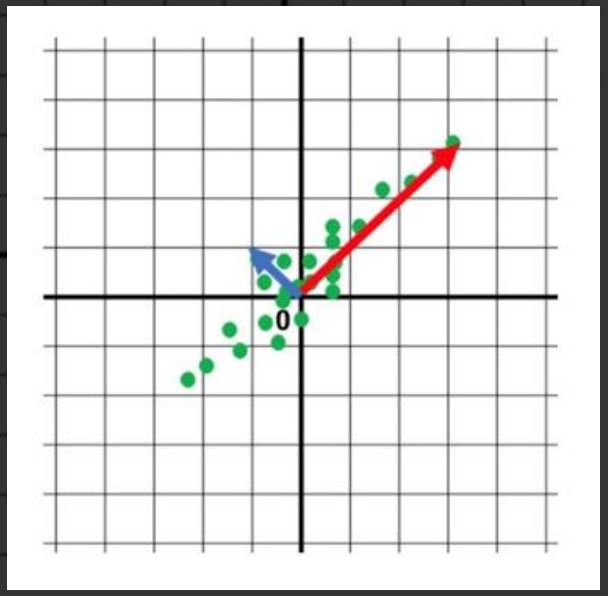
따라서 빨간색 벡터는 공분산 행렬의 고유벡터이며,

이 방향으로 데이터를 projection 시키면 분산을 잘 보존할 수 있음



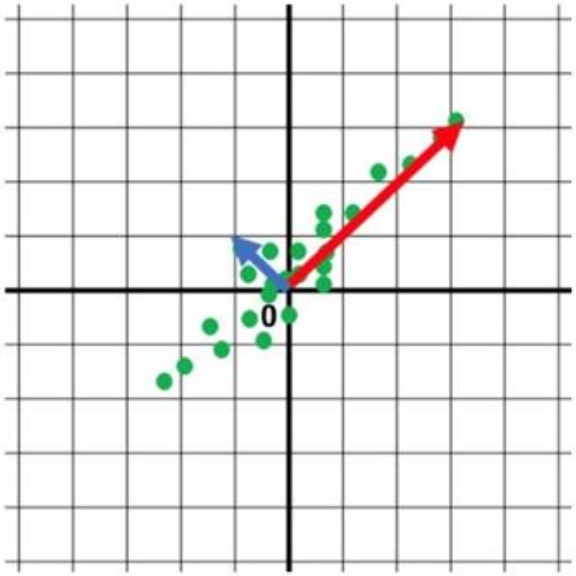
주성분 분석

2차원의 고유벡터는 하나가 아님



따라서 빨간색 벡터가 고유벡터이며
파란색 벡터 또한 공분산 행렬의 고유벡터
이 방향으로 데이터를 projection 시키면 분산을 잘 보존할 수 있음
따라서 어떤 벡터에 projection 해야 할지 판단해야 함

주성분 분석

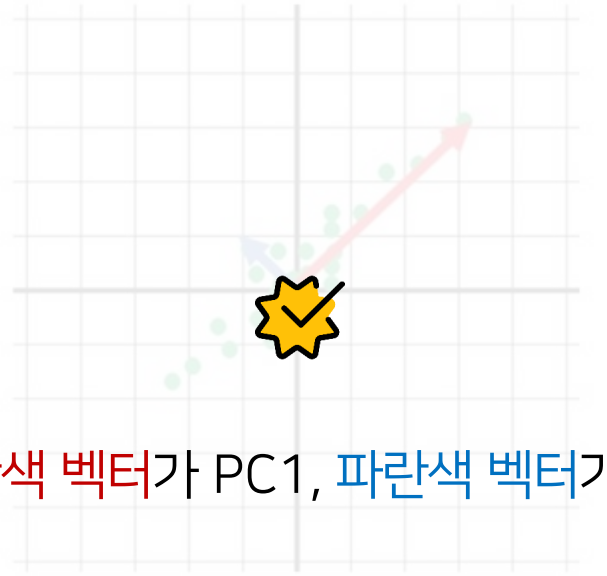


이때 고유값이 쓰이게 됨

고유값은 고유벡터 방향으로 얼마나 데이터가 퍼져있는지를 의미

고유값이 더 큰 빨간색 벡터 방향으로 데이터가 더 많이 분산되어 있음

주성분 분석



따라서 빨간색 벡터가 PC1, 파란색 벡터가 PC2가 됨

이때 고유값이 쓰이게 됨

고유값은 고유벡터 방향으로 얼마나 데이터가 퍼져있는지를 의미

고유값이 더 큰 빨간색 벡터 방향으로 데이터가 더 많이 분산되어 있음

1 주성분 분석

주성분 선택

어떻게, 어디까지 차원을 축소해야 하는지 기준 필요

PCA는 비지도학습이기 때문에

명확한 기준 존재하지 않음

데이터마이닝팀 클린업 1주차 참고

데이터의 분산이 정보량을 말해준다는 것을 이용해

고유값을 기준으로 주성분의 개수를 결정함

d차원의 데이터를 m차원까지 감소시키는 PCA 진행

(d차원의 데이터가 full rank라는 가정 하에)

고유값은 d개이며, $\lambda_1, \lambda_2, \ldots, \lambda_d$ 로 표현 가능

1 주성분 분석

주성분 선택

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{i=1}^{d} \lambda_i} = 0.9$$

전체 데이터 분산 중 90% 정도를

보존하는 것이 좋음

↓

일부 고유값의 합이 전체의 90%가 되는

M개의 변수만을 추출해 차원 축소

(2)

scree plot의 elbow point

↓

elbow point 바로 앞 점까지만 선택

1 주성분 분석

주성분 선택

elbow point

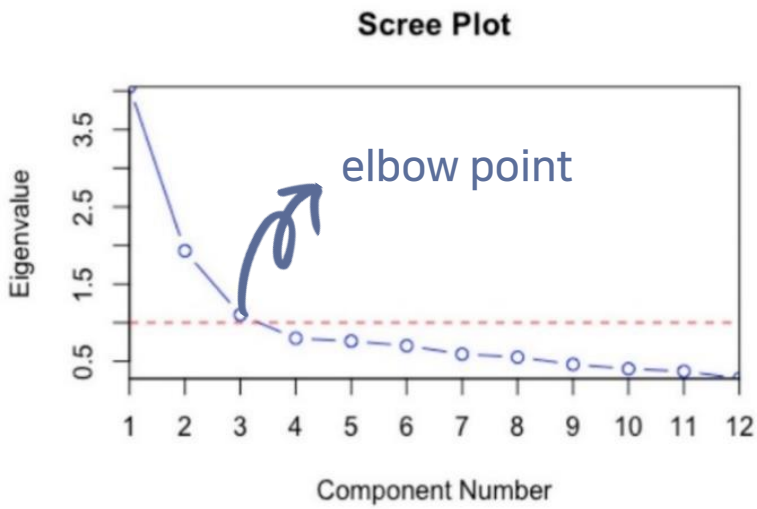
고유값을 기준으로

기울기가 갑자기 변하는 지점

PC3 이후로는 변수의 분산이 작아,

유의한 변수가 되기 힘듦

→ PC2까지만 선택



scree plot의 elbow point

↓

elbow point 바로 앞 점까지만 선택

PCA가 필요한 상황

PCA 과정 요약

데이터를 설명하는 **차원 축을 변경**하고,

그 중에서 **많은 정보를 가진 축만** 남기는 것

변경된 차원의 축은

기존 변수의 조합으로 만들어지기 때문에

축을 해석하기 어려워짐

PCA가 필요한 상황

PCA 과정 요약

데이터를 설명하는 **차원 축을 변경**하고,

그 중에서 **많은 정보를 가진 축만** 남기는 것

언제 PCA를 사용해야 할까?

변경된 차원의 축은

기존 변수의 조합으로 만들어지기 때문에

을 해석하기 어려워짐

높은 다중공선성

Multicollinearity

시각화

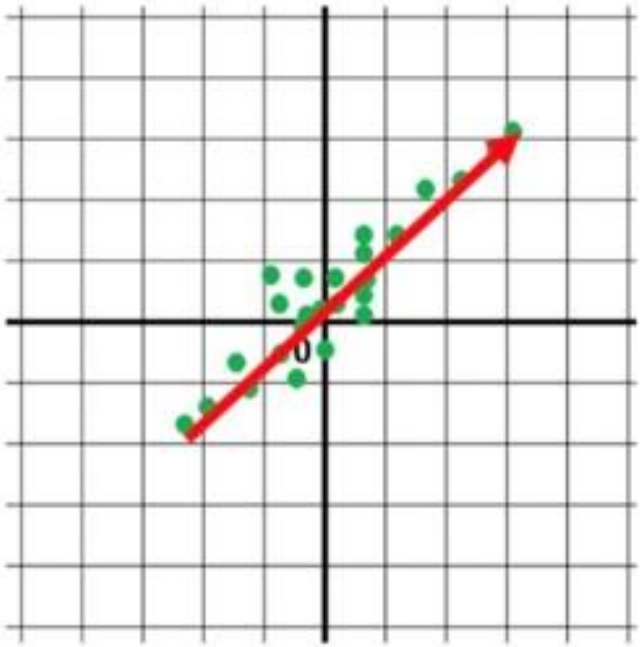
Visualization

1 주성분분석

PCA가 필요한 상황

1. 높은 다중공선성

Ex)



x 변수 간 상관관계가 높다면, PCA 고려

빨간색 벡터 한 축으로 데이터를 표현해도

소실되는 정보량이 적기에 PCA 사용 적합

다중공선성 문제가 심각할수록 잘 작동함

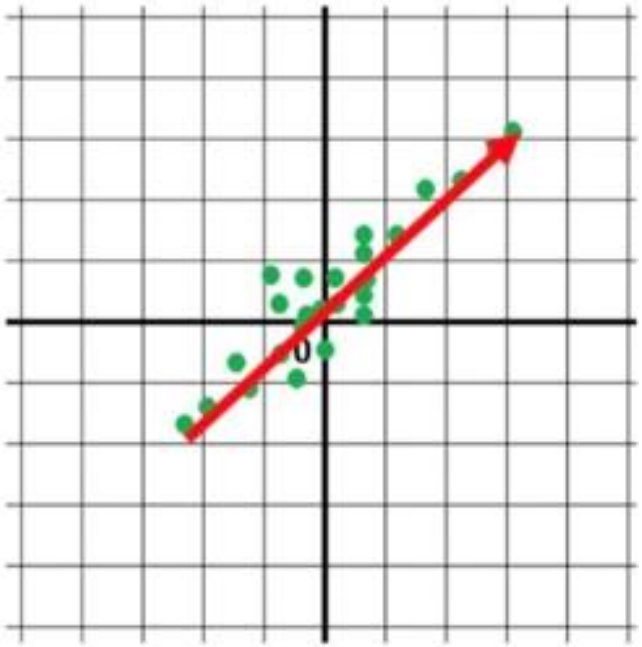
But, 상관관계가 높은 변수 중 중요도가
떨어지는 변수를 제거하는
변수선택법도 고려할 필요가 있음

1 주성분분석

PCA가 필요한 상황

1. 높은 다중공선성

Ex)



x 변수 간 상관관계가 높다면, PCA 고려

빨간색 벡터 한 축으로 데이터를 표현해도

소실되는 정보량이 적기에 PCA 사용 적합

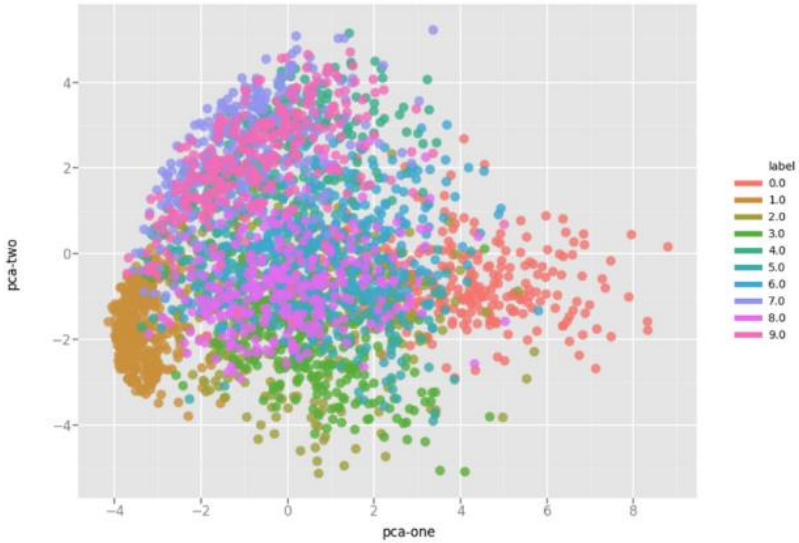
다중공선성 문제가 심각할수록 잘 작동함

But, 상관관계가 높은 변수 중 중요도가
떨어지는 변수를 제거하는
변수선택법도 고려할 필요가 있음

PCA가 필요한 상황

2. 시각화

Ex) PCA 클러스터링 시각화



고차원 데이터 클러스터링은 시각화 어려움

변수 간 선형 연관성이 크다면,

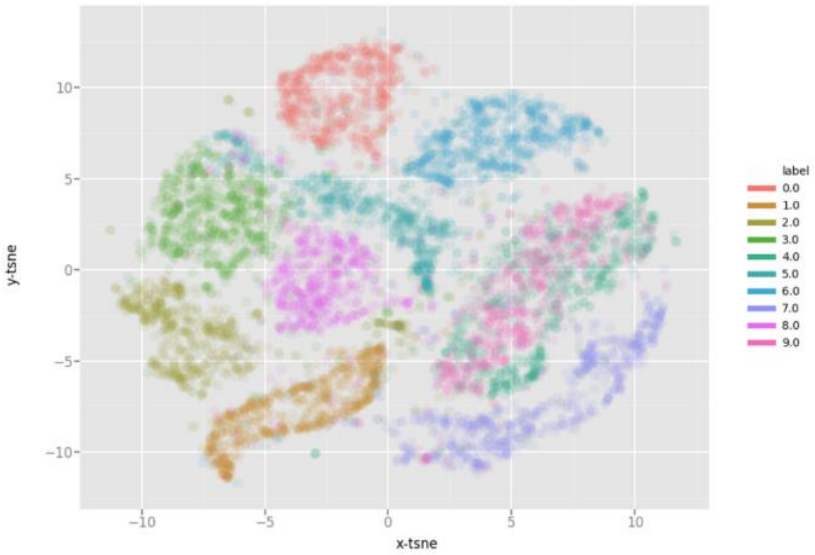
PCA로 차원을 축소할 수 있음

2차원이나 3차원으로 시각화 가능

PCA가 필요한 상황

2. 시각화

Ex) t-SNE 클러스터링 시각화



일반적인 고차원 데이터 시각화 방법

t-SNE나 UMAP 같은 알고리즘 사용

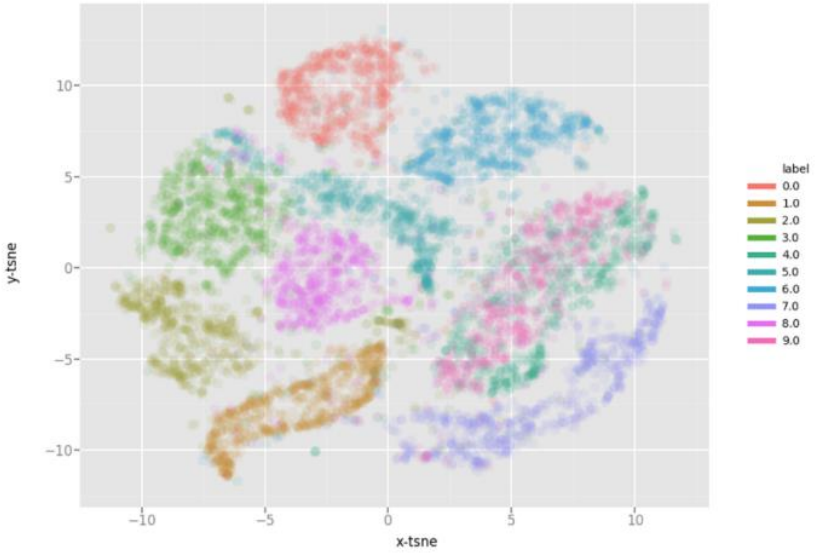
이웃 간의 거리를 잘 유지하며 데이터 축소

시각화할 때 더 명확하게 데이터를 구분해줌

PCA가 필요한 상황

2. 시각화

Ex) t-SNE 클러스터링 시각화



하지만 t-SNE의 경우,

① 데이터의 특성이나 매핑되는 위치가 변해 **시각화에만 사용될 뿐, 새로운 x변수로 활용 X**

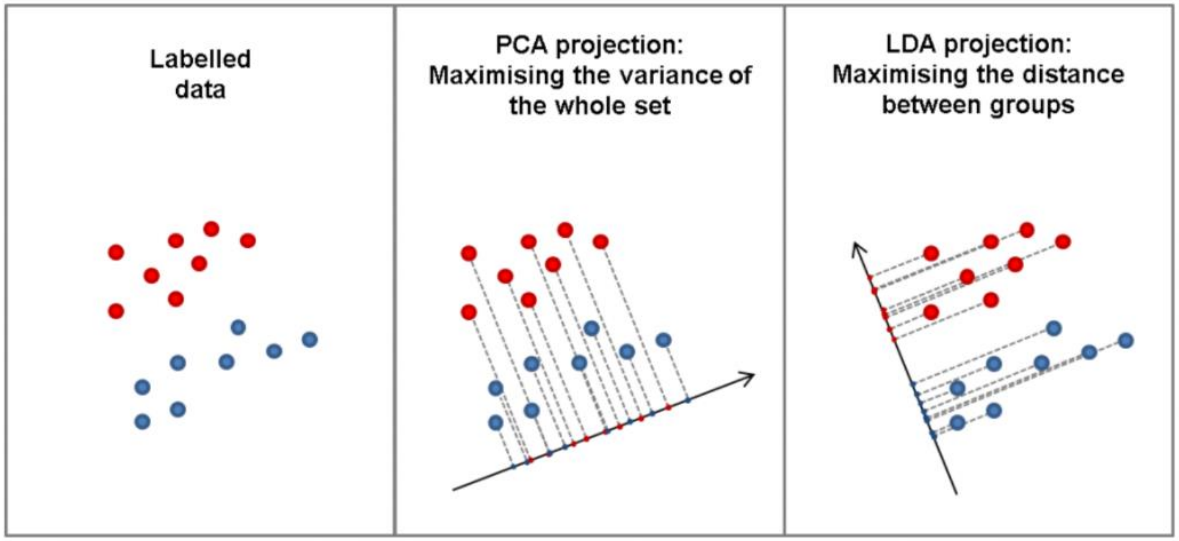
② 많은 계산이 요구되기 때문에 **PCA를 통해 차원을 축소한 후 사용되기도 함**

데이터의 특성, 분석 목적에 맞게 알고리즘 선택!

그 외의 차원축소 기법

LDA (Linear Discriminant Analysis)

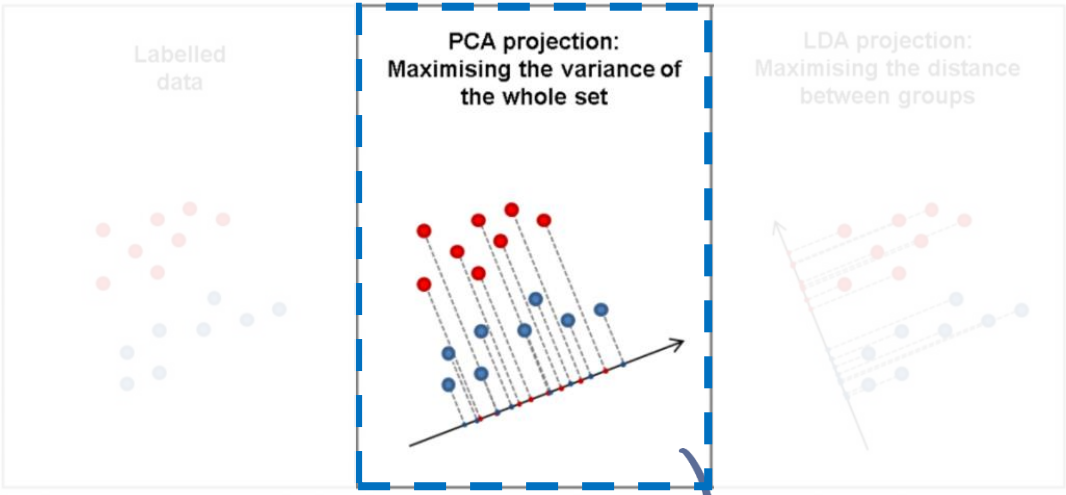
PCA와 달리 **클래스 분류에 용이한 차원 축소 방법**



그 외의 차원축소 기법

LDA (Linear Discriminant Analysis)

PCA와 달리 **클래스 분류에 용이한 차원 축소 방법**



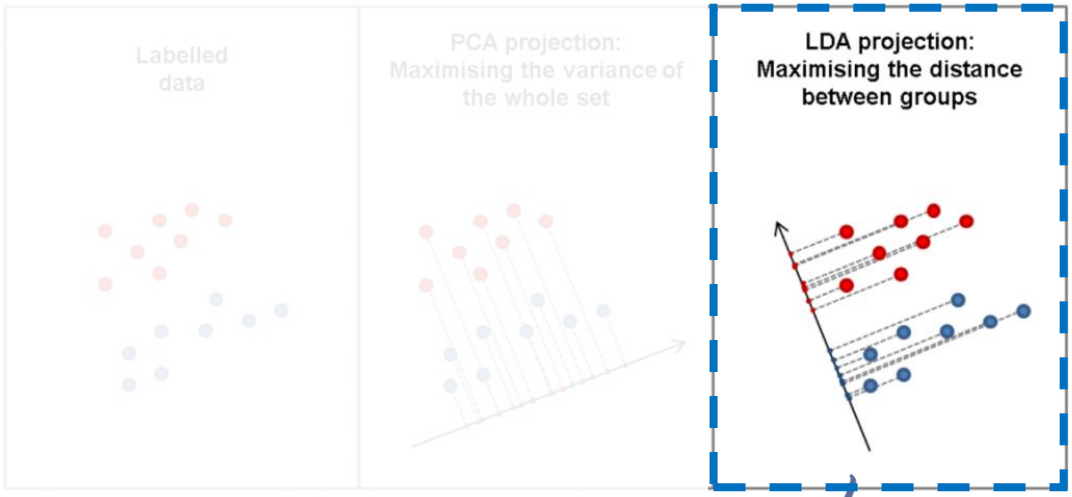
PCA로 축소

클래스와 상관없이 마구잡이로 projection

그 외의 차원축소 기법

LDA (Linear Discriminant Analysis)

PCA와 달리 **클래스 분류에 용이한 차원 축소 방법**



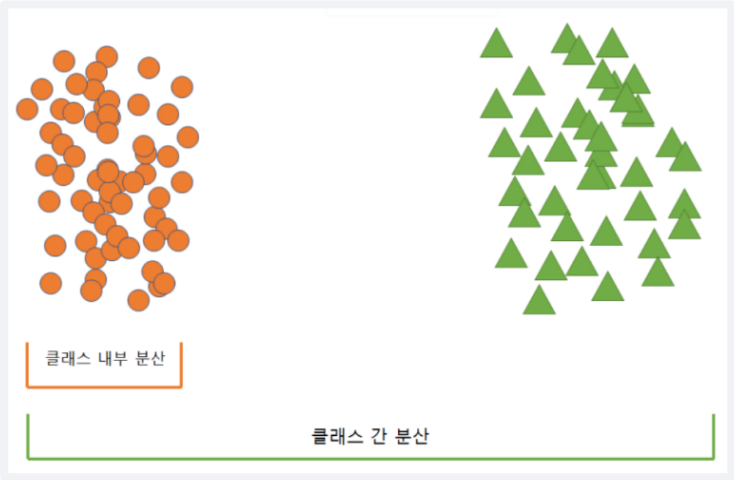
LDA로 축소

클래스가 잘 구분되어 projection

그 외의 차원축소 기법

LDA (Linear Discriminant Analysis)

PCA와 달리 **클래스 분류에 용이한 차원 축소 방법**



LDA는 **클래스 내부에서의 분산은 작게**,
클래스 간 분산은 크게 만드는 축을 찾아 축소

그 외의 차원축소 기법

LDA (Linear Discriminant Analysis)

PCA와 달리 **클래스 분류에 용이한 차원 축소 방법**

차원 축소 방법

PCA

공분산 행렬의
고유 벡터와 고유값을 구함

LDA

클래스 내부 분산 및 클래스 간
분산 행렬의 고유벡터와 고유값을 구함

그 외의 차원축소 기법

LDA (Linear Discriminant Analysis)

PCA와 달리 **클래스 분류에 용이한 차원 축소 방법**

중요한 차이는

개별 클래스를 분별할 수 있는 기준을

최대한 유지하면서 차원을 축소한다는 것

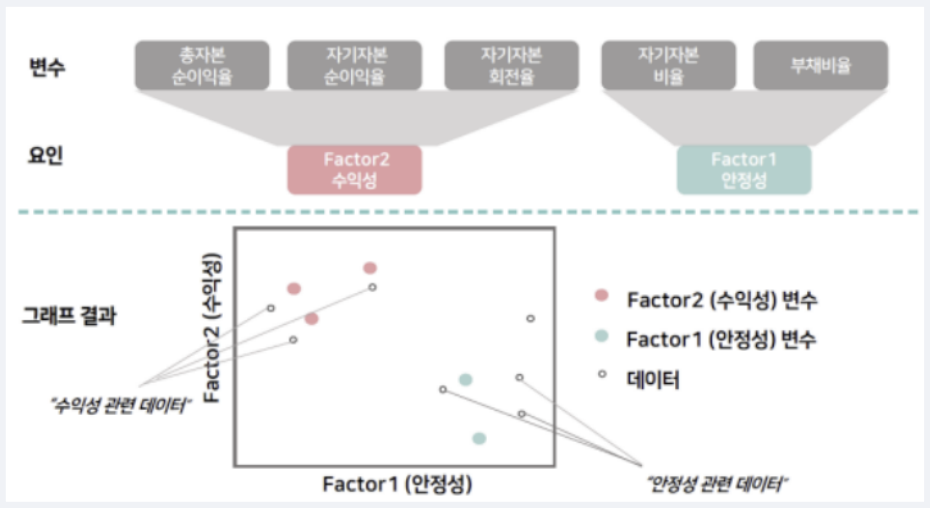
1 주성분 분석

그 외의 차원축소 기법

요인분석 (Factor Analysis)

변수들의 상관관계를 고려하여 **내재된 요인을 추출**해내 **요인별로 변수를 묶어줌**

변수들이 어떤 요인의 영향을 받는가를 알아보는 것



Ex)

자본 관련 **변수 5개가 있을 때,**

'안정성'과 '수익성'이라는

2가지 요인으로 묶어 요인을 파악함

그 외의 차원축소 기법

요인분석 (Factor Analysis)

변수들의 상관관계를 고려하여 **내재된 요인을 추출**해내 **요인별로 변수를 묶어줌**

변수들이 어떤 요인의 영향을 받는가를 알아보는 것

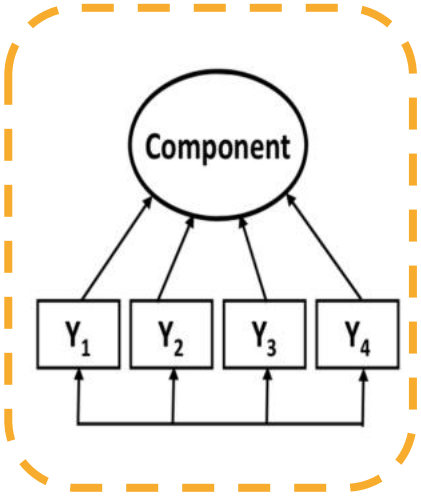
PCA, 요인분석 모두 **고차원의 변수를 줄이는 방법**이며,

공분산 행렬의 고유값, 고유벡터로 구할 수 있다는 점에서 비슷함

하지만, 두 모델의 사용 목적은 완전히 다름

1 주성분 분석

요인분석의 목적



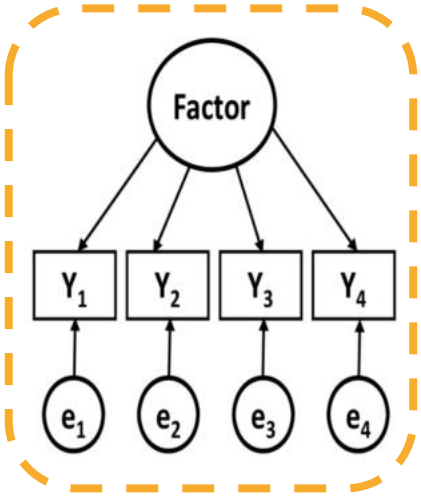
PCA

모든 독립변수를 조합해 **주성분**을 만들어내고

분산 대부분을 설명할 수 있는 선에서 일부를 선택

1 주성분 분석

요인분석의 목적



요인분석

데이터의 여러 변수들이 가지고 있는 공분산 구조를 밝히는 것

즉, **변수의 의미 및 특성 파악**이 우선 목표

1 주성분 분석

그 외의 차원축소 기법

요인분석

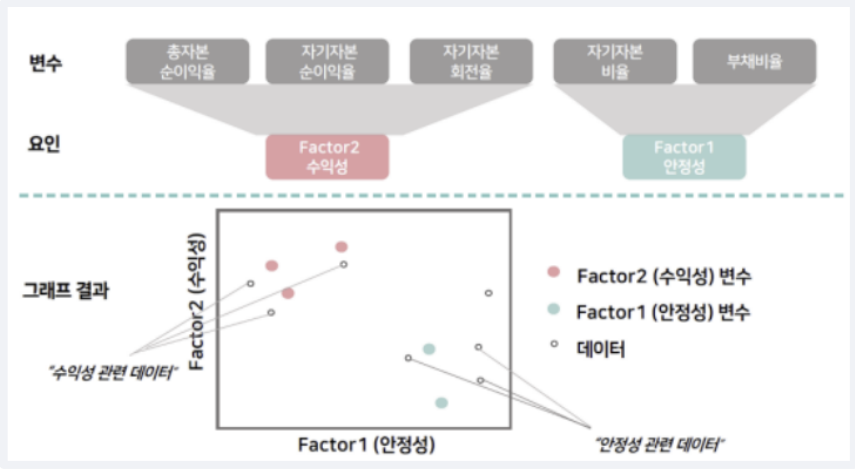
탐색적 요인분석(EFA)

요인을 어림해 만듦

확인적 요인분석(CFA)

요인모형이 적합하게

만들어졌는지 확인함



Ex) EFA

어떤 요인이 있는지 모르는 상황에서

요인이 2개 정도 나오는 것을 확인하고

각각 안정성과 수익성이라고 이름 붙임

1 주성분 분석

그 외의 차원축소 기법

요인분석

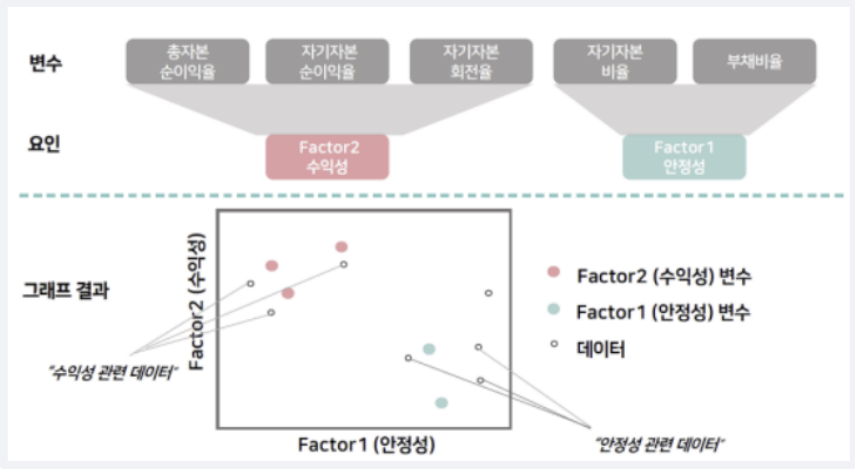
탐색적 요인분석(EFA)

요인을 어림해 만듦

확인적 요인분석(CFA)

요인모형이 적합하게

만들어졌는지 확인함



Ex) CFA

안정성과 수익성이라는

두 요인이 있다는 가설을 세운 후

요인분석을 통해 요인 모형 검정

2

특이값 분해 응용

2 특이값 분해 응용

SVD 기반 PCA

고유값 분해를 통한 PCA

데이터의 공분산 행렬을 구해야 함

데이터가 클수록 공분산 행렬을 구하기 위한 계산량 증가!

특이값 분해 사용!

공분산 행렬을 메모리에 저장할 필요 X → 효율적

2 특이값 분해 응용

SVD 기반 PCA

고유값 분해를 통한 PCA

데이터의 공분산 행렬을 구해야 함

데이터가 클수록 공분산 행렬을 구하기 위한 계산량 증가!

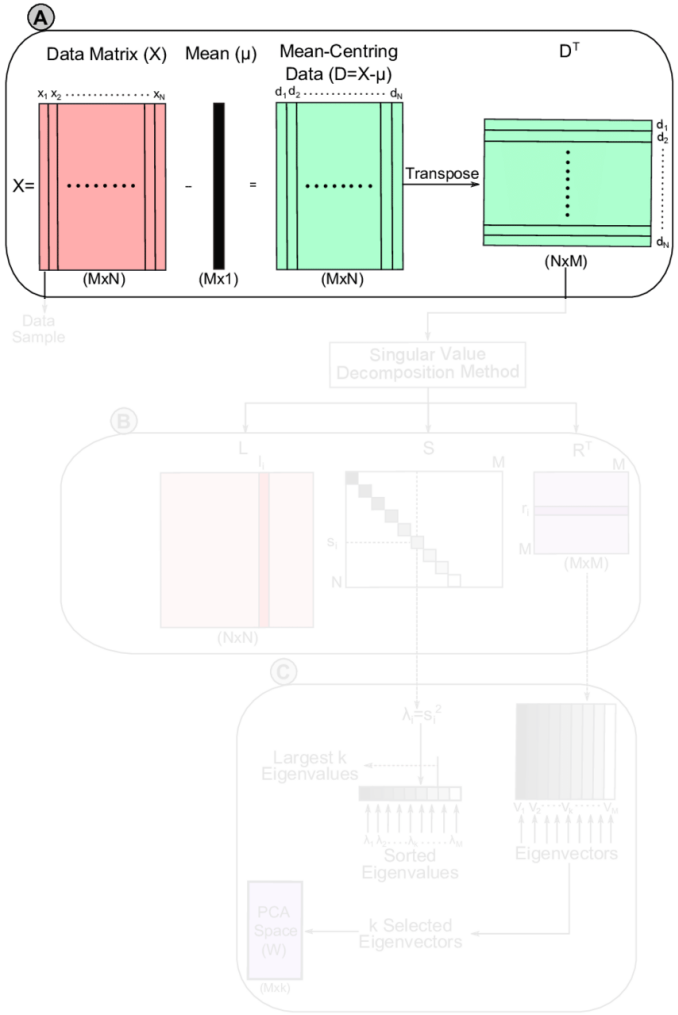
특이값 분해 사용!

공분산 행렬을 메모리에 저장할 필요 X → 효율적

2 특이값 분해 응용

SVD 기반 PCA

- X 표준화 (scale 영향 제거)
- $X = U\Sigma V^T$를 찾기 위해 SVD 사용
- 주성분에 데이터 project ($XV = U\Sigma$)
  V의 일부 열만 선택해 차원 축소



2 특이값 분해 응용

SVD 기반 PCA

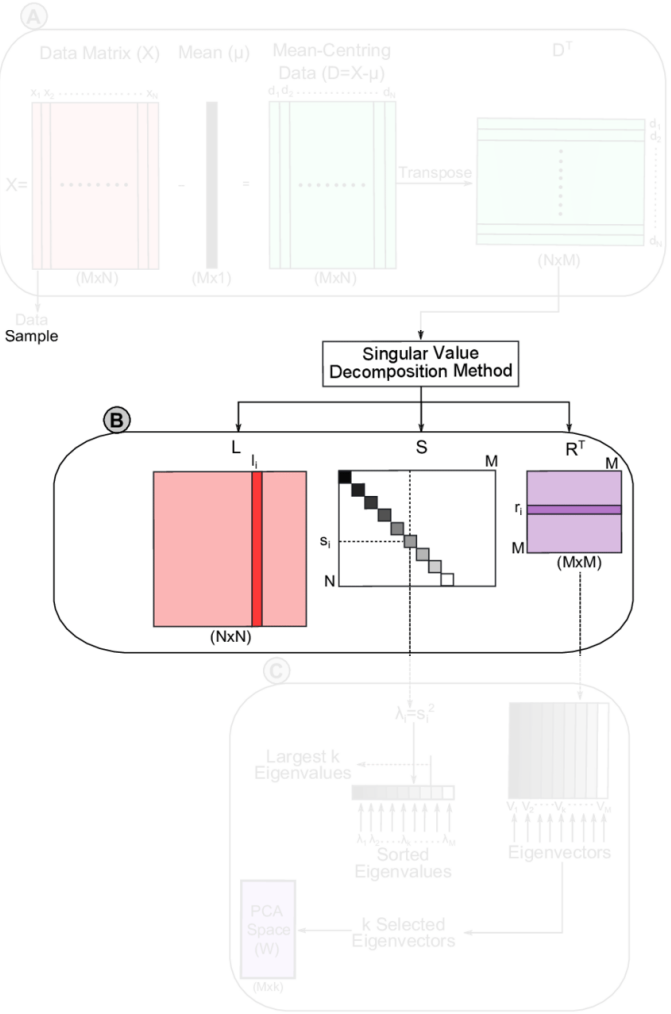
X 표준화 (scale 영향 제거)

$X = U\Sigma V^T$를 찾기 위해 SVD 사용

V^T: $X^T X$의 고유벡터 행렬 = 주성분

주성분에 데이터 project ($XV = U\Sigma$)

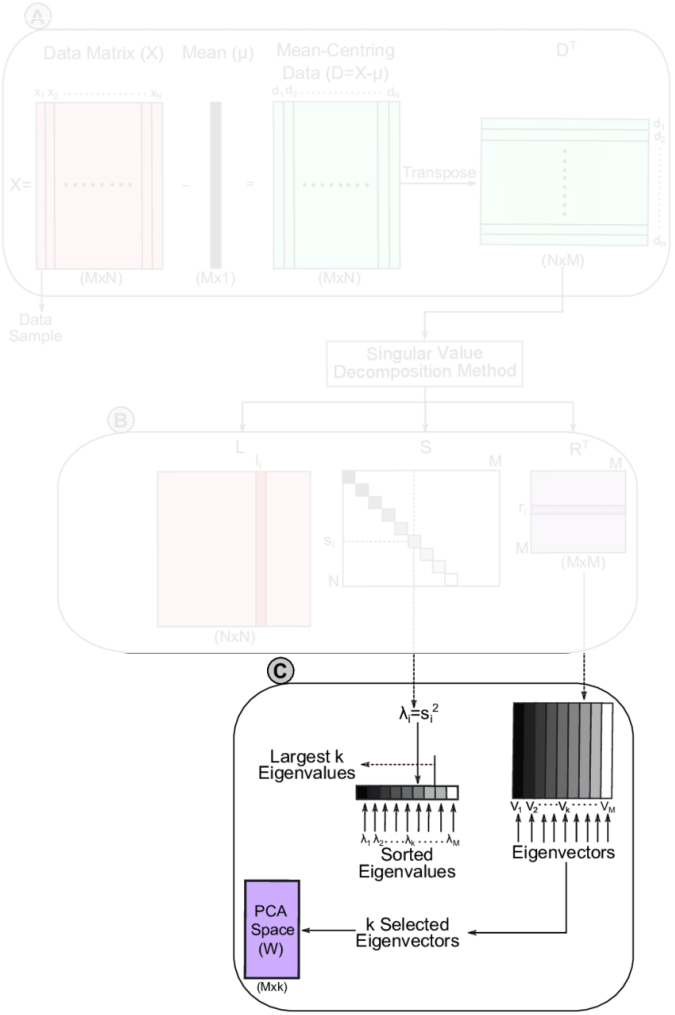
V의 일부 열만 선택해 차원 축소



2 특이값 분해 응용

SVD 기반 PCA

- X 표준화 (scale 영향 제거)
- $X = U\Sigma V^T$를 찾기 위해 SVD 사용
- 주성분에 데이터 project ($XV = U\Sigma$)
  V의 일부 열만 선택해 차원 축소



2 특이값 분해 응용

잠재의미분석(LSA)

토픽모델링

전체 문서의 주제를 연구자가 지정한 개수만큼 압축,

각 문서들이 어떤 주제를 가지는지 확인

LSA는 토픽모델링의 시초로 Truncated SVD를 사용

문서

문서 1: Pizza

문서 2: Pizza Hamburger Cookie

문서 3: Hamburger

문서 4: Ramen

문서 5: Sushi

문서 6: Ramen Sushi

	문서1	문서2	문서3	문서4	문서5	문서6
Pizza	1	1	0	0	0	0
Hamburger	0	1	1	0	0	0
Cookie	0	1	0	0	0	0
Ramen	0	0	0	1	0	1
Sushi	0	0	0	0	1	1

2 특이값 분해 응용

잠재의미분석(LSA)

토픽모델링

전체 문서의 주제를 연구자가 지정한 개수만큼 압축,

각 문서들이 어떤 주제를 가지는지 확인

LSA는 토픽모델링의 시초로 Truncated SVD를 사용

문서

문서 1: Pizza

문서 2: Pizza Hamburger Cookie

문서 3: Hamburger

문서 4: Ramen

문서 5: Sushi

문서 6: Ramen Sushi

	문서1	문서2	문서3	문서4	문서5	문서6
Pizza	1	1	0	0	0	0
Hamburger	0	1	1	0	0	0
Cookie	0	1	0	0	0	0
Ramen	0	0	0	1	0	1
Sushi	0	0	0	0	1	1

2 특이값 분해 응용

잠재의미분석(LSA)

토픽모델링

전체 문서의 주제를 연구자가 지정한 개수만큼 압축,

각 문서들이 어떤 주제를 가지는지 확인

LSA는 토픽모델링의 시초로 Truncated SVD를 사용

문서

문서 1: Pizza

문서 2: Pizza Hamburger Cookie

문서 3: Hamburger

문서 4: Ramen

문서 5: Sushi

문서 6: Ramen Sushi

문서와 단어 간의 관계에

어떤 topic이 내재되어 있다고 가정

	문서1	문서2	문서3	문서4	문서5	문서6
Pizza	1	1	0	0	0	0
Hamburger	0	1	1	0	0	0
Cookie	0	1	0	0	0	0
Ramen	0	0	0	1	0	1
Sushi	0	0	0	0	1	1

특이값 분해를 통해 찾고자 함

2 특이값 분해 응용

잠재의미분석(LSA)

단어-문서 행렬 A

	문서1	문서2	문서3	문서4	문서5	문서6
Pizza	1	1	0	0	0	0
Hamburger	0	1	1	0	0	0
Cookie	0	1	0	0	0	0
Ramen	0	0	0	1	0	1
Sushi	0	0	0	0	1	1

=

U

	T1	T2	T3	T4	T5
W1	0.6	0	0	0.7	-0.3
W2	0.6	0	0	-0.7	-0.3
W3	0.5	0	0	0	0.9
W4	0	0.7	-0.7	0	0
W5	0	0.7	0.7	0	0

×

Σ

	T1	T2	T3	T4	T5	T6
T1	1.9	0	0	0	0	0
T2	0	1.7	0	0	0	0
T3	0	0	1	0	0	0
T4	0	0	0	1	0	0
T5	0	0	0	0	0.5	0

×

V^T

	D1	D2	D3	D4	D5	D6
T1	0.3	0.9	0.3	0	0	0
T2	0	0	0	0.4	0.4	0.8
T3	0	0	0	-0.7	0.7	0
T4	0.7	0	-0.7	0	0	0
T5	-0.6	0.5	-0.6	0	0	0

4 고유값과 고유벡터

잠재의미분석(LSA)

$A = U\Sigma V^T$ 연산은 Term X Document의 관계를 다음과 같이 표현

$$Term \times Document$$

$$= (Topic \times Term)\ (Topic \times Topic)\ (Document \times Topic)$$

U

	T1	T2	T3	T4	T5
W1	0.6	0	0	0.7	-0.3
W2	0.6	0	0	-0.7	-0.3
W3	0.5	0	0	0	0.9
W4	0	0.7	-0.7	0	0
W5	0	0.7	0.7	0	0

단어와 토픽의 관계

×

Σ

	T1	T2	T3	T4	T5	T6
T1	1.9	0	0	0	0	0
T2	0	1.7	0	0	0	0
T3	0	0	1	0	0	0
T4	0	0	0	1	0	0
T5	0	0	0	0	0.5	0

토픽의 영향력

×

V^T

	D1	D2	D3	D4	D5	D6
T1	0.3	0.9	0.3	0	0	0
T2	0	0	0	0.4	0.4	0.8
T3	0	0	0	-0.7	0.7	0
T4	0.7	0	-0.7	0	0	0
T5	-0.6	0.5	-0.6	0	0	0

토픽과 문서의 관계

4 고유값과 고유벡터

잠재의미분석(LSA)

$A = U\Sigma V^T$ 연산은 Term X Document의 관계를 다음과 같이 표현

문서당 하나씩 나타나고 있는 주제들을 압축

영향력이 큰 순으로 2개의 토픽 값만을 사용하여 Truncated SVD 진행

U

	T1	T2	T3	T4	T5
W1	0.6	0	0	0.7	-0.3
W2	0.6	0	0	-0.7	-0.3
W3	0.5	0	0	0	0.9
W4	0	0.7	-0.7	0	0
W5	0	0.7	0.7	0	0

단어와 토픽의 관계

×

Σ

	T1	T2	T3	T4	T5	T6
T1	1.9	0	0	0	0	0
T2	0	1.7	0	0	0	0
T3	0	0	1	0	0	0
T4	0	0	0	1	0	0
T5	0	0	0	0	0.5	0

토픽의 영향력

×

V^T

	D1	D2	D3	D4	D5	D6
T1	0.3	0.9	0.3	0	0	0
T2	0	0	0	0.4	0.4	0.8
T3	0	0	0	-0.7	0.7	0
T4	0.7	0	-0.7	0	0	0
T5	-0.6	0.5	-0.6	0	0	0

토픽과 문서의 관계

4 고유값과 고유벡터

잠재의미분석(LSA)

원래 행렬 A

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Truncated SVD로 만든 유사행렬 A'

$$\begin{bmatrix} 0.342 & 1.026 & 0.342 & 0 & 0 & 0 \\ 0.342 & 1.026 & 0.342 & 0 & 0 & 0 \\ 0.285 & 0.855 & 0.285 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.476 & 0.476 & 0.952 \\ 0 & 0 & 0 & 0.476 & 0.476 & 0.952 \end{bmatrix}$$

오차는 있지만 A의 경향성을 유지하고 있음

V^T 를 통해 **잠재되어 있는 주제**가 무엇인지,

행을 통해 **토픽에 대한 각 단어의 영향력** 확인 가능

4 고유값과 고유벡터

잠재의미분석(LSA)

토픽의 영향력을 계산하여 정리하면 (ΣV^T),

Σ

	T1	T2
T1	1.9	0
T2	0	1.7

$\times$

V^T

	D1	D2	D3	D4	D5	D6
T1	0.3	0.9	0.3	0	0	0
T2	0	0	0	0.4	0.4	0.8

=

	D1	D2	D3	D4	D5	D6
T1	0.57	1.71	0.57	0	0	0
T2	0	0	0	0.68	0.68	1.36

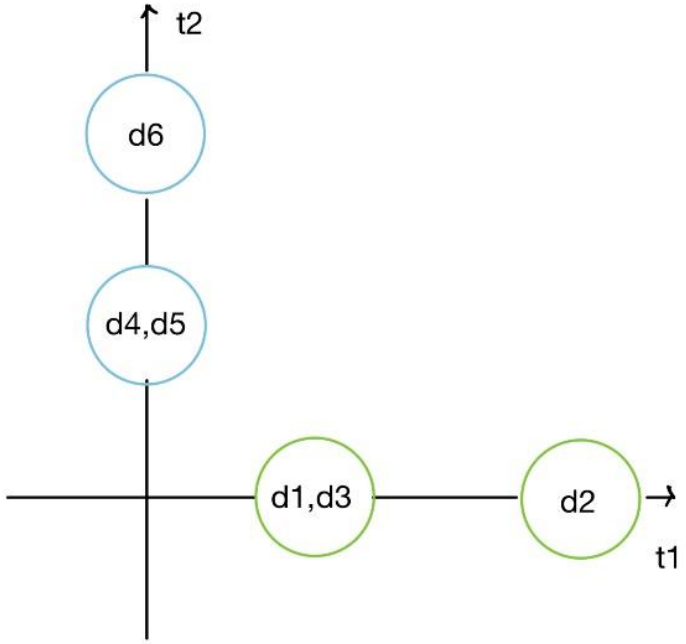
4 고유값과 고유벡터

잠재의미분석(LSA)

토픽의 영향력을 계산하여 정리하면(ΣV^T) ...

	D1	D2	D3	D4	D5	D6
T1	0.57	1.71	0.57	0	0	0
T2	0	0	0	0.68	0.68	1.36

D1, D2, D3 → T1(양식)
D4, D5, D6 → T2(일식)



4 고유값과 고유벡터

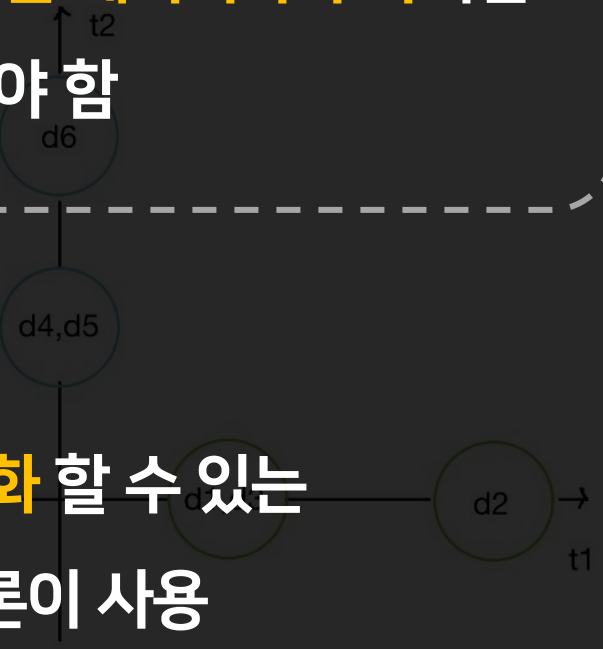
잠재의미분석(LSA)

토픽의 영향력을 계산하여 정리하면(ΣV^T) ...

SVD 특성 상 이미 계산된 LSA에 **새로운 데이터가 추가**되면

처음부터 다시 계산해야 함

	D1	D2	D3	D4	D5	D6
T1	0.57	1.71	0.57	0	0	0
T2	0	0	0	0.68	0.68	1.36



최근에는 **단어의 의미를 벡터화** 할 수 있는

인공 신경망 기반의 방법론이 사용

D1, D2, D3 → T1(양식)
D4, D5, D6 → T2(날씨)

자연어 처리에 대한 자세한 내용은 3주차 딥러닝팀 교안 참고!

3

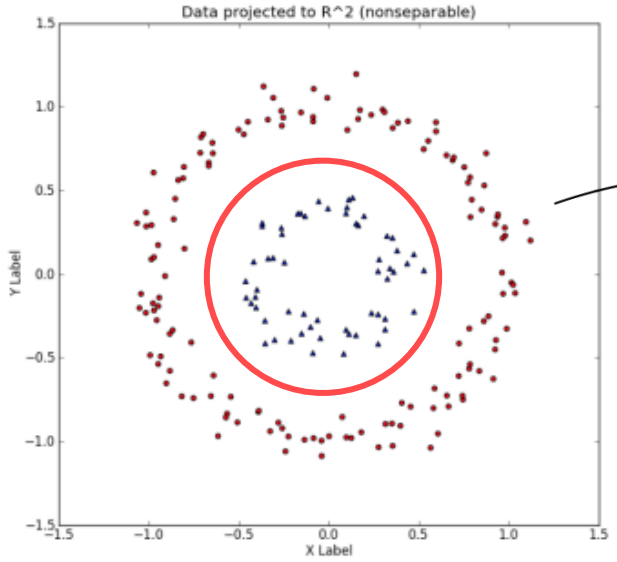
커널과 커널 트릭

3 커널과 커널 트릭

커널 (kernel)

더 높은 차원에서 데이터를 바라보는 것

즉 **데이터를 높은 차원으로 이동시켜 그 공간에서 분류**하는 것



분류를 위해

비선형 모델 필요

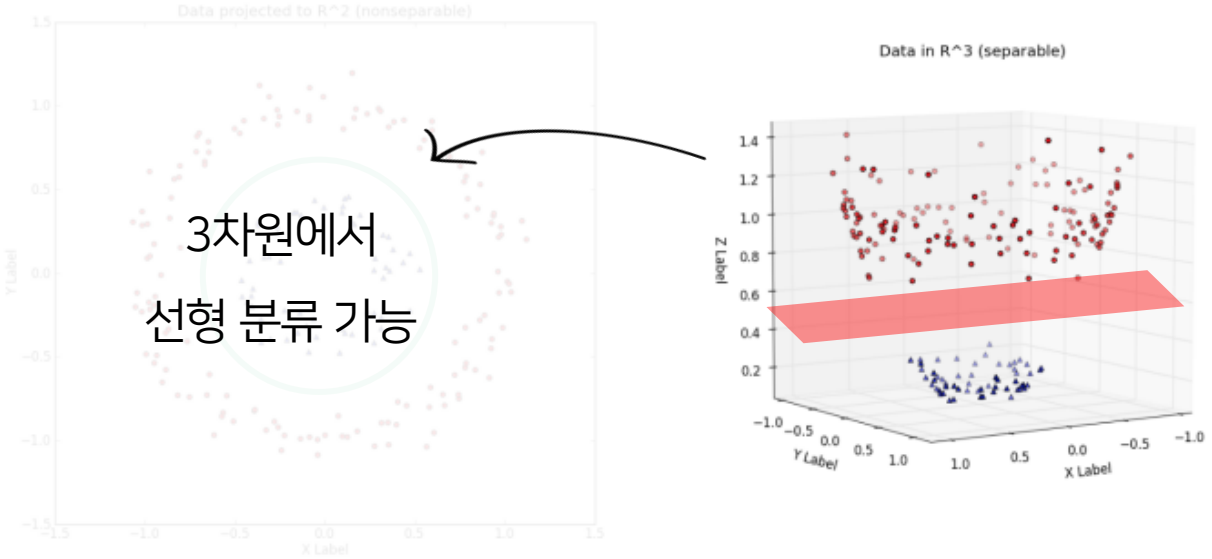
선형분류 가능

3 커널과 커널 트릭

커널 (kernel)

더 높은 차원에서 데이터를 바라보는 것

즉 **데이터를 높은 차원으로 이동시켜 그 공간에서 분류**하는 것



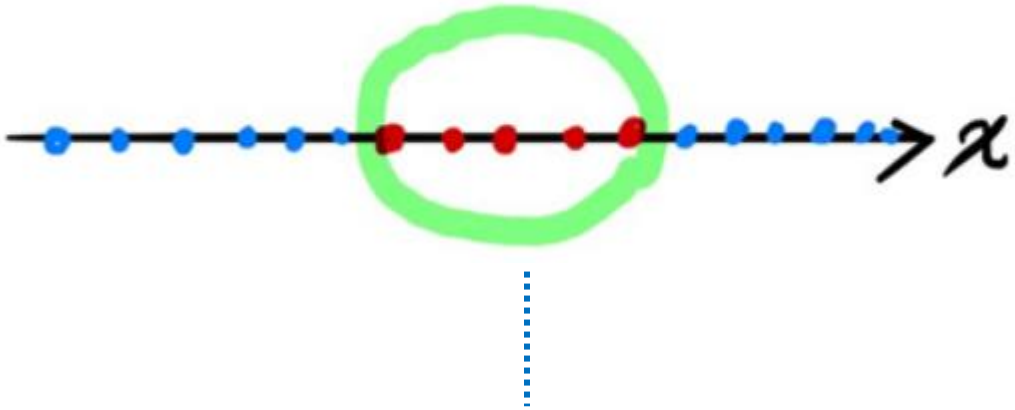
3 커널과 커널 트릭

커널 (kernel)

기저함수 (Basis function)

Input space X에 존재하는 데이터들을

feature space로 옮겨주는 mapping

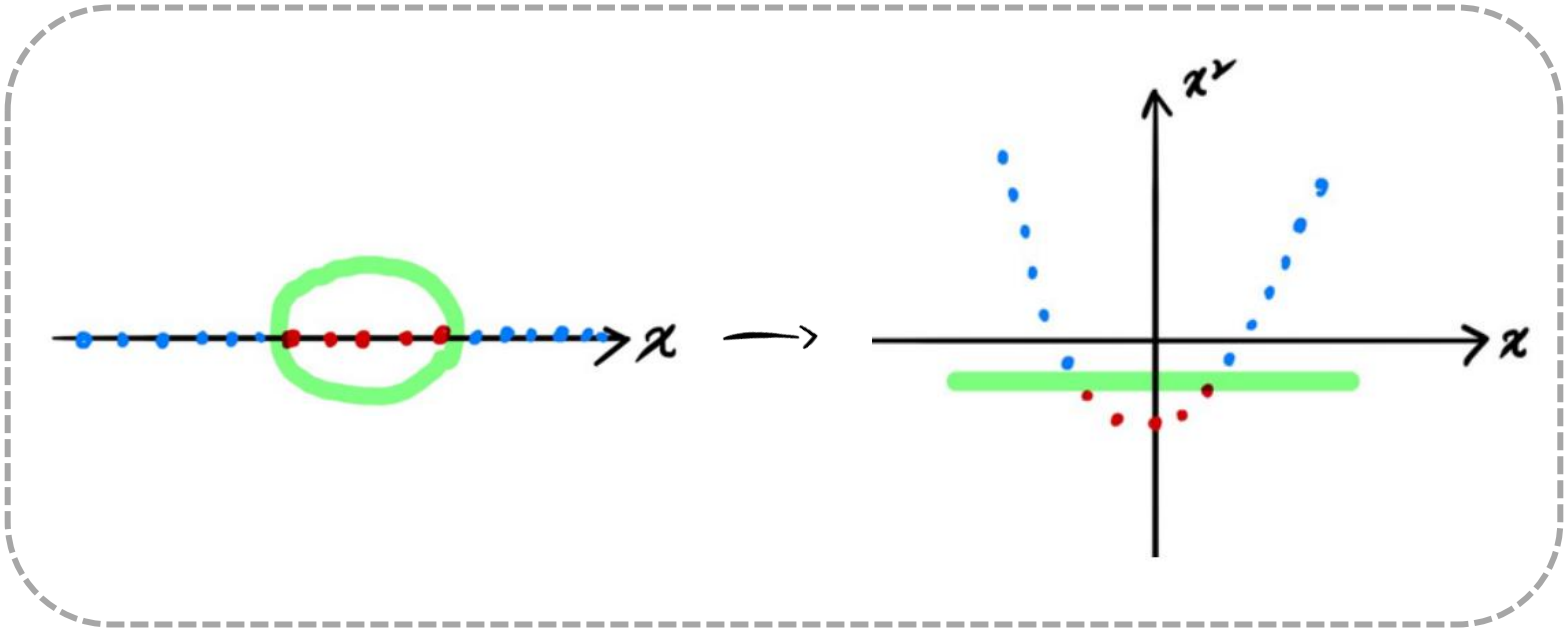


선형적인 방법으로는 두 범주를 분류할 수 없음

3 커널과 커널 트릭

커널 (kernel)

Input space $x \rightarrow \{x, x^2\}$로 변경



선형문제로 변환

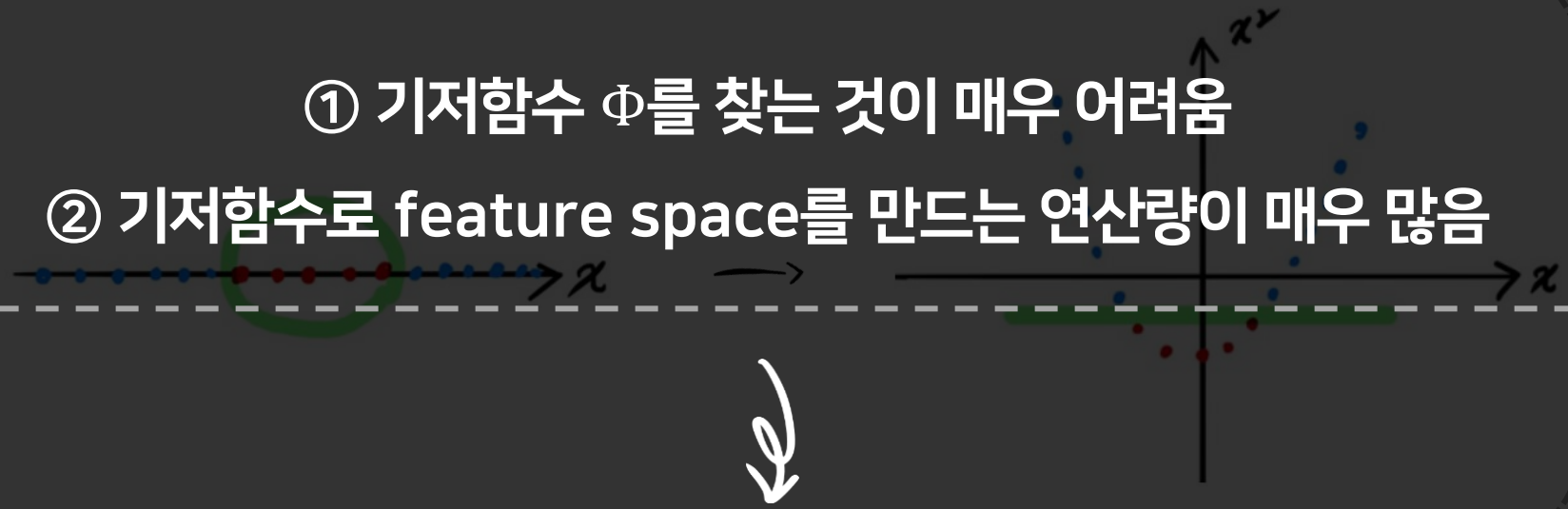
3 커널과 커널 트릭

커널 (kernel)

기저함수 사용의 문제점

① 기저함수 Φ를 찾는 것이 매우 어려움

② 기저함수로 feature space를 만드는 연산량이 매우 많음



커널 트릭(kernel trick) 사용

선형문제로 변환

커널 (kernel)

커널 (kernel)

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle, \qquad \forall x \in X$$

커널은 어떻게 구해야 할까?

실제 고차원으로의 mapping에 대응되는 커널이 존재한다는 것을 어떻게 알 수 있을까?

커널 (kernel)

커널 (kernel)

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle, \qquad \forall x \in X$$

커널은 어떻게 구해야 할까?

실제 고차원으로의 mapping에 대응되는 커널이 존재한다는 것을 어떻게 알 수 있을까?

3 커널과 커널 트릭

커널 (kernel)

커널 (kernel)

Mercer's theorem

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle, \forall x \in X$$

① 커널 k가 스칼라를 출력하는 연속함수

② 커널 k로 구성된 행렬이 대칭행렬이고 positive semi-definite

(대각원소가 0초과)

두 조건을 만족하는 커널에 대하여

$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$를 만족하는 mapping Φ가 존재함

3 커널과 커널 트릭

커널 트릭 (kernel trick)

Mercer's Theorem을 **만족**하는 임의의 함수 k는 **모두 커널**로 사용 가능

⋮

기저함수 Φ를 찾지 않고

$\Phi(x_i)^T\Phi(x_j)$를 $K(x_i, x_j)$**로 대체**해주면 빠르게 계산 가능

SVM, KPCA같이 **고차원에서의 변수간 유사도**를 사용할 때

기저함수 대신 **커널로 내적값을 구하는 것** = 커널 트릭

3 커널과 커널 트릭

커널 트릭 (kernel trick)

Mercer's Theorem을 **만족**하는 임의의 함수 k는 **모두 커널**로 사용 가능

⋮

기저함수 Φ를 찾지 않고

$\Phi(x_i)^T\Phi(x_j)$를 $K(x_i, x_j)$**로 대체**해주면 빠르게 계산 가능

SVM, KPCA같이 **고차원에서의 변수간 유사도**를 사용할 때

기저함수 대신 커널로 내적값을 구하는 것 = 커널 트릭

Kernel functions : polynomial kernel

Polynomial kernel

$$K(x, y) = (x^T y + r)^d$$

r = polynimial의 계수, d = polynomial의 차수

종속변수와 **독립변수**의 관계가 polynomial function에 의해 잘 설명되는 경우 사용

Ex)

$$k(a, b) = \left(\mathrm{a} \cdot \mathrm{b} + \frac{1}{2}\right)^2$$

커널 이용

$$= \left(\mathrm{a} \times \mathrm{b} + \frac{1}{2}\right)\left(\mathrm{a} \times \mathrm{b} + \frac{1}{2}\right) = \mathrm{ab} + \mathrm{a}^2\mathrm{b}^2 + \frac{1}{4}$$

기저함수 이용

$$= \left(\mathrm{a}, \mathrm{a}^2, \frac{1}{2}\right) \cdot \left(\mathrm{b}, \mathrm{b}^2, \frac{1}{2}\right)$$

$r = {}^1/_2$, $d = 2$

Kernel functions : polynomial kernel

Polynomial kernel

$$K(x, y) = (x^T y + r)^d$$

r = polynimial의 계수, d = polynomial의 차수

종속변수와 **독립변수**의 관계가 polynomial function에 의해 잘 설명되는 경우 사용

Ex)

$$k(a, b) = \left(\mathrm{a} \cdot \mathrm{b} + \frac{1}{2}\right)^2$$

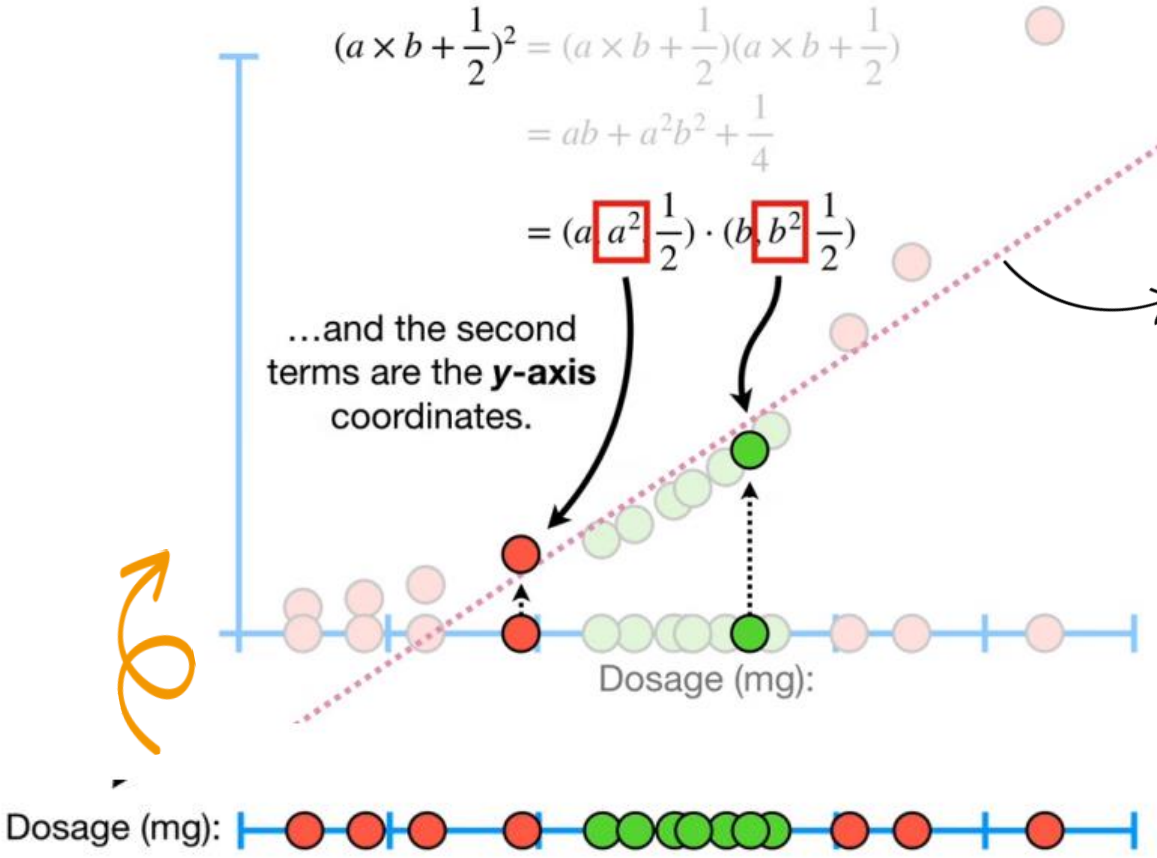
$$= \left(\mathrm{a} \times \mathrm{b} + \frac{1}{2}\right)\left(\mathrm{a} \times \mathrm{b} + \frac{1}{2}\right) = \mathrm{ab} + \mathrm{a}^2\mathrm{b}^2 + \frac{1}{4}$$

(X값, y값, z값) ← $= \left(\mathrm{a}, \mathrm{a}^2, \frac{1}{2}\right) \cdot \left(\mathrm{b}, \mathrm{b}^2, \frac{1}{2}\right)$

$r = {}^1/_2$, $d = 2$

3 커널과 커널 트릭

Kernel functions : polynomial kernel



고차원으로 mapping 했을 때 linear하게 분류 가능

3 커널과 커널 트릭

Kernel functions : polynomial kernel

기저함수 대신 커널 함수를 사용해

두 점 a=4, b=9의 내적을 쉽게 구할 수 있음

$$k(4,9) = \left(4 \cdot 9 + \frac{1}{2}\right)^2 = (36.5)^2 = 1332.25$$

3 커널과 커널 트릭

Kernel functions : RBF kernel (Gaussian kernel)

RBF kernel

$$K(x, y) = e^{-\gamma(x-y)^2}$$

γ = 두 점 사이의 거리 $(x - y)^2$를 조절해주는 하이퍼 파라미터

두 점 사이의 **거리**를 정함으로써 **다른 점에 얼마나 영향**을 줄 것인지 결정

차원이 무한한 특성공간에 mapping하는 것

관측값에 대한 **사전정보가 없을 때** 주로 사용

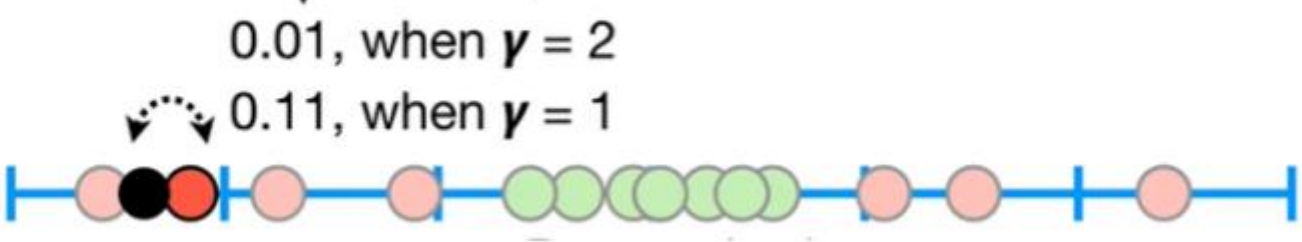
3 커널과 커널 트릭

Kernel functions : RBF kernel (Gaussian kernel)

고차원에서의 두 관측값의 관계를 나타냄

$e^{-\gamma(a-b)^2}$

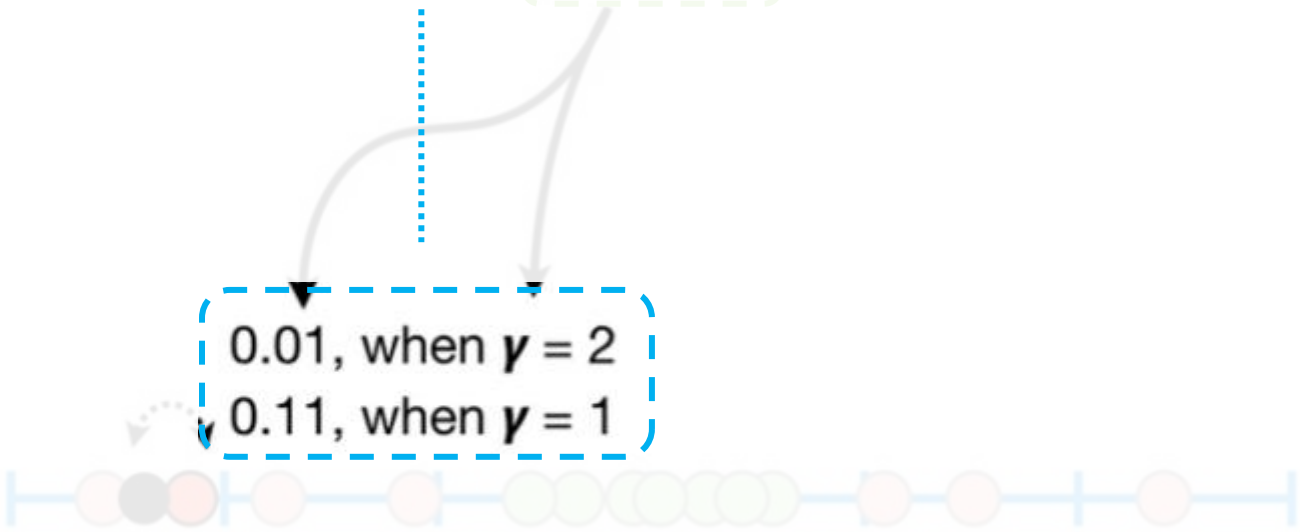
고차원으로 mapping한 이후
두 벡터의 내적값



Kernel functions : RBF kernel (Gaussian kernel)

γ가 작으면 점 사이의 거리를 좁히게 되어 영향력 증가

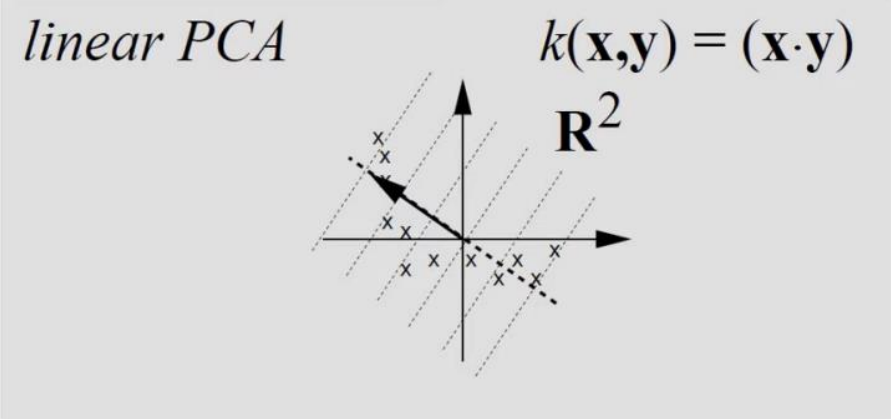
γ가 크면 점 사이의 거리를 늘리게 되어 영향력 감소



γ로 관측값이 다른 관측값에 주는 영향 스케일링

커널 활용 – Kernel PCA

PCA : 데이터의 **분산을 최대한 보존**하며 차원 축소



일반적인 PCA를 사용했을 때 **주성분은 선형**

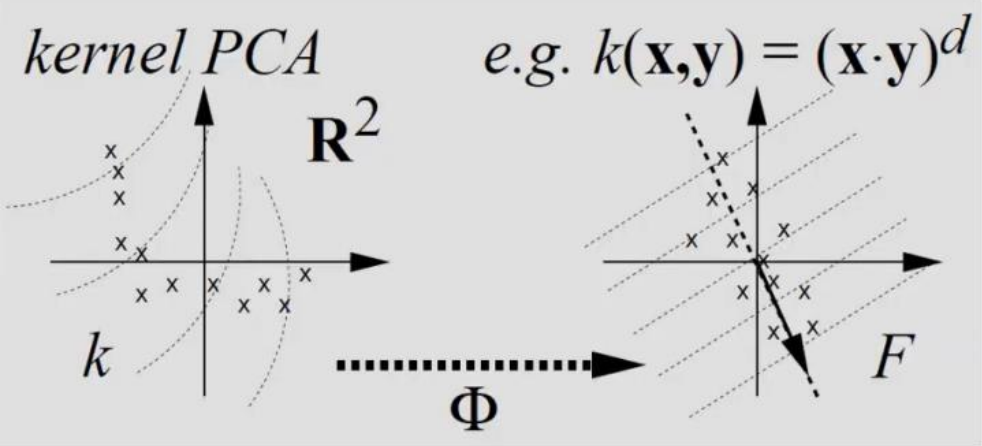
하지만 **실제데이터는 곡선 형태**로 분포해 있음

⋮

데이터에 맞지 않는 주성분으로 투영하는 문제 발생

커널 활용 – Kernel PCA

커널을 사용해 고차원으로 mapping한 후 PCA 진행



데이터에 맞지 않는 주성분으로 투영하던 기존 문제 해결

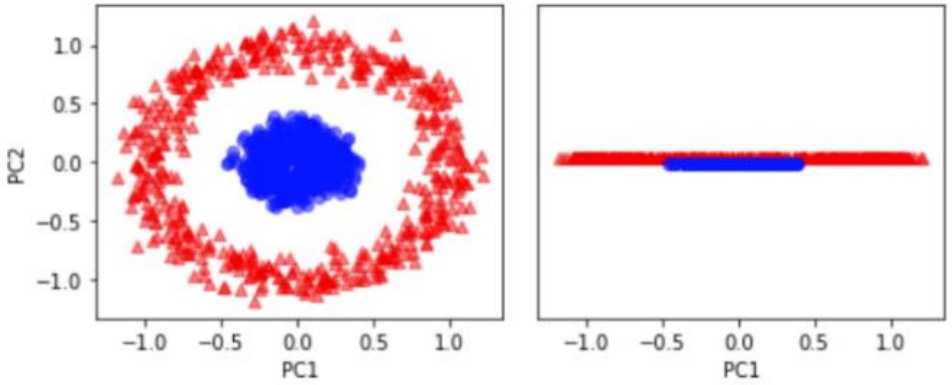
커널 활용 – Kernel PCA

Kernel PCA

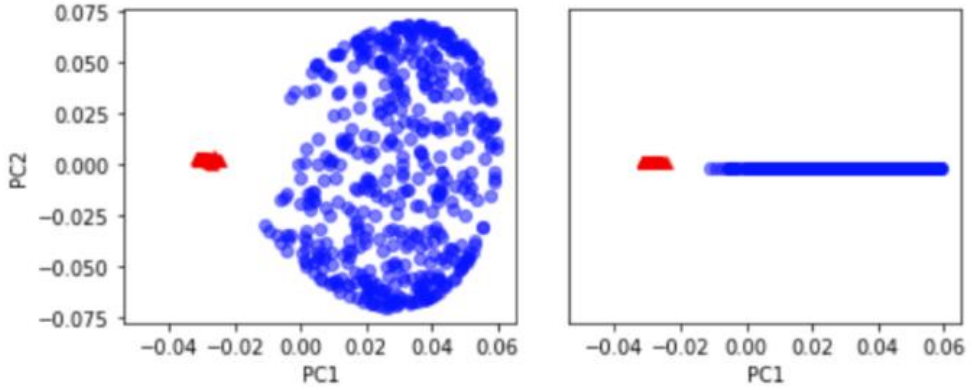
비선형 구조 데이터의 차원을 더 잘 축소할 수 있는 방법

커널 활용 – Kernel PCA

일반적인 PCA를 사용했을 경우



RBF Kernel PCA를 사용했을 경우



3 커널과 커널 트릭

커널 활용 – Kernel PCA

선형분류가 가능한 고차원의 특성을 이용해

차원을 축소했을 때(RBF Kernel PCA) 시각화가 잘 됨

Kernel PCA는 컴퓨터 비전, 음성 인식 등

비선형적 구조를 가진 데이터를 **시각화**해서

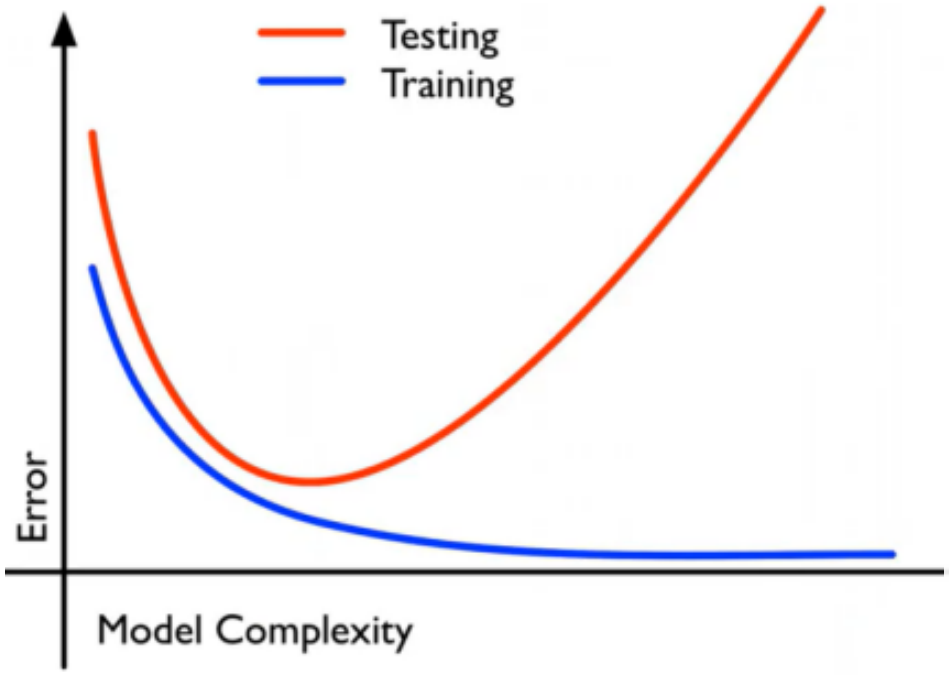
군집, 트렌드, 이상치 등을 파악하는데 사용

커널 활용 – Kernel PCA

Kernel PCA를 사용했을 때 노이즈를 훨씬 효과적으로 줄인 것을 확인할 수 있음

3 커널과 커널 트릭

커널과 딥러닝



모델이 복잡해질수록
Training error는 줄지만
Training data에 과적합되어
Test error가 점차 증가함

↓

매우 복잡한 모델인
딥러닝에 대한
관심이 적었음

3 커널과 커널 트릭

커널과 딥러닝

매우 복잡한 딥러닝 모델이 현실에서 잘 작동

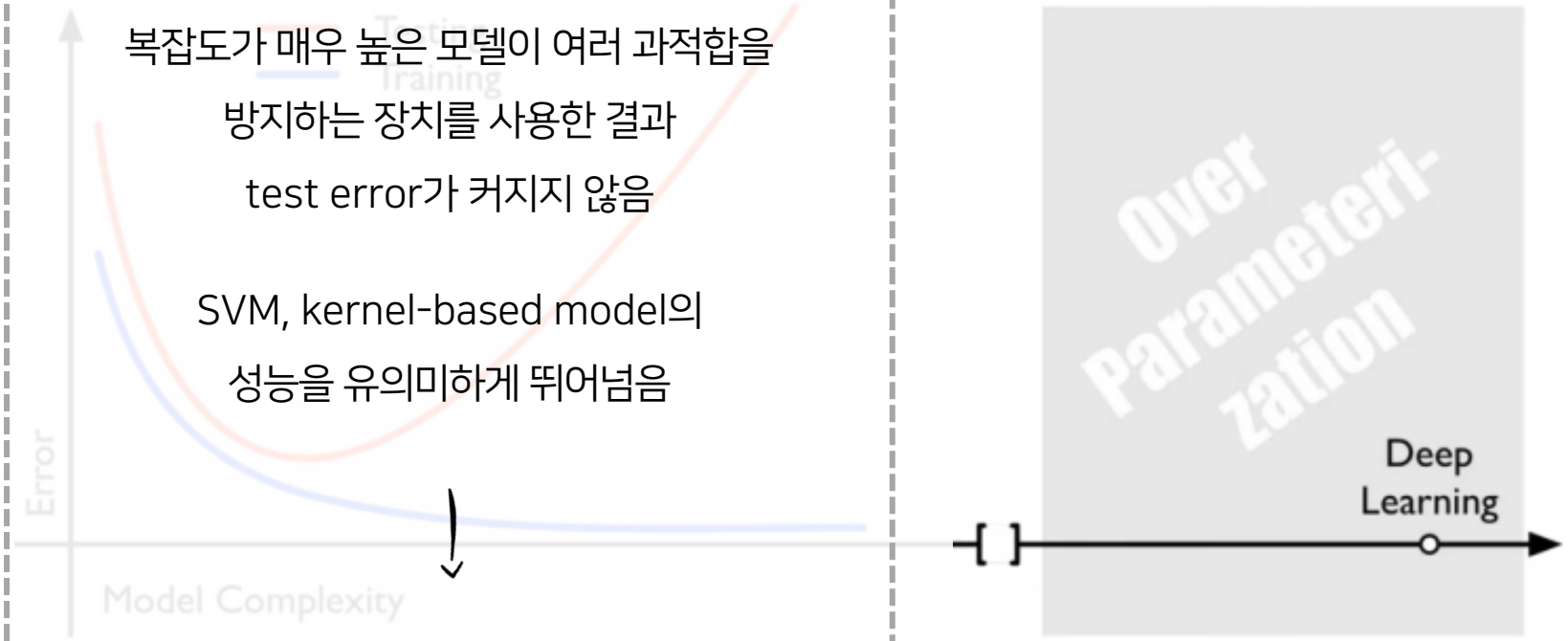
복잡도가 매우 높은 모델이 여러 과적합을

방지하는 장치를 사용한 결과

test error가 커지지 않음

SVM, kernel-based model의

성능을 유의미하게 뛰어넘음



딥러닝에 대한 관심 증가

3 커널과 커널 트릭

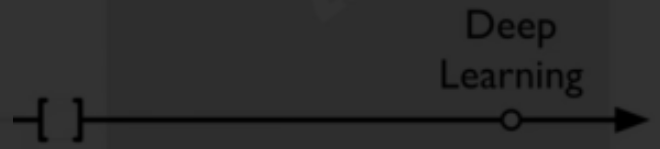
커널과 딥러닝

전통적인 통계모델을 알 필요가 없을까?

여전히 전통적인 머신 러닝이 잘 작동하는 분야가 있으므로

최대한 많은 것을 아는 것이 중요

- ✓ 분포가정을 통한 구간 추정
- ✓ 변수 간 관계 파악



데이터와 분석 목적에 맞는 모델을 사용하는 것이 중요

감사합니다

클린업 탈출 ~~ 기저 ~~

-범주팀 익명의 그녀-

결국 모든 것은 기저를 어떻게 바꾸는 가와 관련이 있기에 기저는 선대에서 매우 중요한 개념입니다.
또한 각각 독립적이지만 하나라도 없으면 그 공간을 생성하지 못하기에 모든 존재가 필요하다는
아주 주요한 의미를 내포하고 있다고 생각합니다 ㅎㅎ기저선대 파이팅~
-선대팀장님-

누가
버렸을까….

죄송합니다…