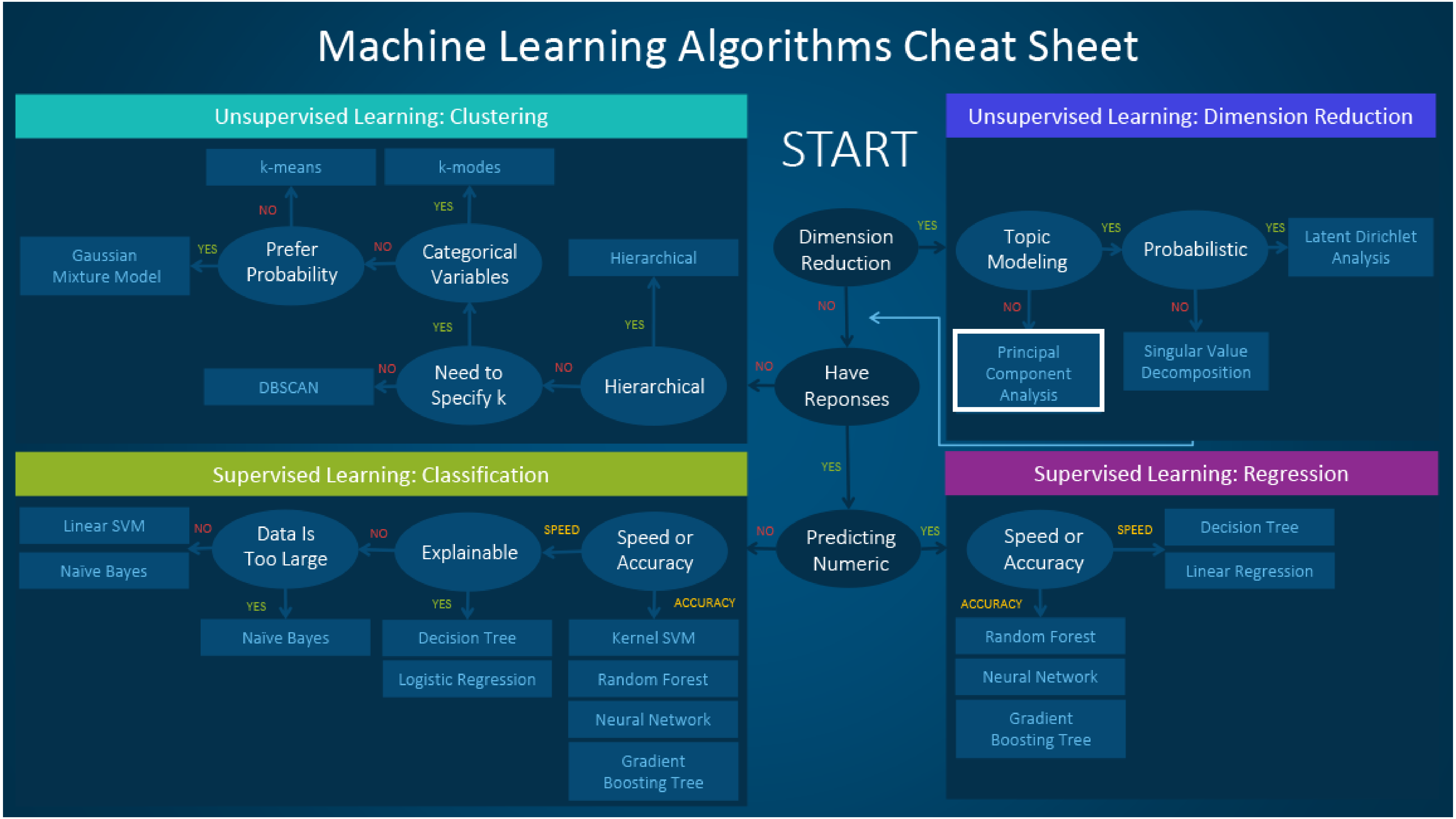
R 교육 세미나

ToBig's 8기 김민정

Principal Component Analysis

주성분 분석

Contents



Machine Learning Algorithms Cheat Sheet
Unsupervised Learning: Clustering
Unsupervised Learning: Dimension Reduction
START
k-means
k-modes
NO
YES
Gaussian Mixture Model
Prefer Probability
Categorical Variables
Hierarchical
Dimension Reduction
Topic Modeling
Probabilistic
Latent Dirichlet Analysis
DBSCAN
Need to Specify k
Hierarchical
Have Reponses
Principal Component Analysis
Singular Value Decomposition
Supervised Learning: Classification
Supervised Learning: Regression
Linear SVM
Naïve Bayes
Data Is Too Large
Explainable
SPEED
Speed or Accuracy
Predicting Numeric
Speed or Accuracy
Decision Tree
Linear Regression
ACCURACY
Naïve Bayes
Decision Tree
Logistic Regression
Kernel SVM
Random Forest
Neural Network
Gradient Boosting Tree
Random Forest
Neural Network
Gradient Boosting Tree

Unit 01 | Intro

주성분 분석?

변수들의 선형결합을 통해 전체 정보를 최대한 설명할 수 있는
서로 독립적인 인공변수들을 유도하여 해석하는 다변량 분석방법

Unit 01 | Intro

주성분 분석?

변수들의 선형결합을 통해 전체 정보를 최대한 설명할 수 있는
서로 독립적인 인공변수들을 유도하여 해석하는 다변량 분석방법

➡ 여러 변수의 정보를 담고있는 주성분이라는 새로운 변수를 생성(Feature Extraction)해 차원을 축소(Dimension Reduction)하는 기법

➡ Data를 scale해주고 고차원 데이터 중 중요한 차원을 골라주어
학습 시 수렴 속도와 성능을 개선시켜 전처리과정에서 많이 쓰인다고 함

Unit 01 | Intro

Idea

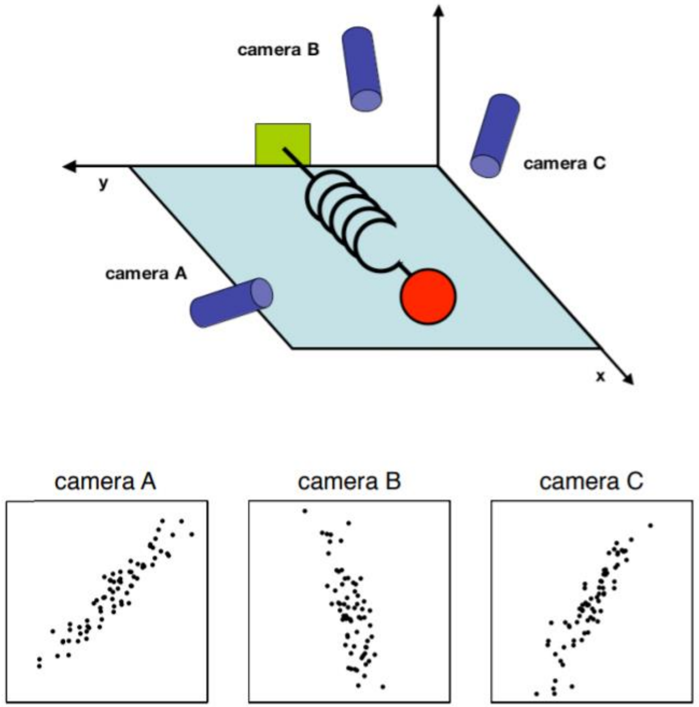


FIG. 1 A toy example. The position of a ball attached to an oscillating spring is recorded using three cameras A, B and C. The position of the ball tracked by each camera is depicted in each panel below.

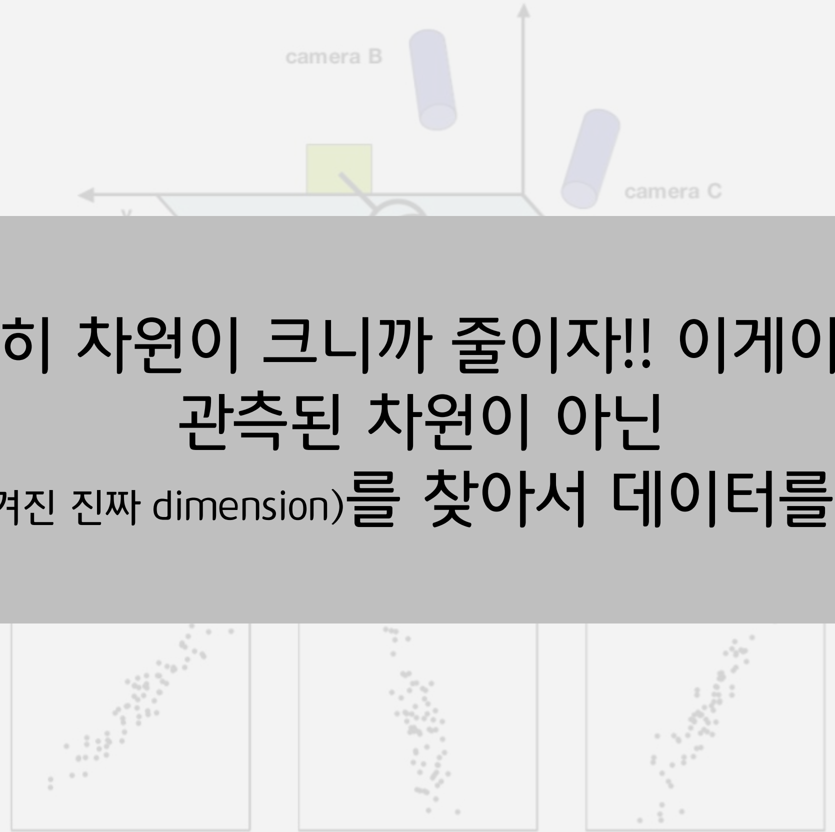
스프링 운동(kx)은 변수 하나로 설명가능
즉, 이 data를 만드는 근본 변수는 하나 !

3대의 카메라 ⇒ 3차원?
20대의 카메라 ⇒ 20차원?
80대 ⇒ 80차원?
800대 ⇒ 800차원???

"A Tutorial on PCA", J. Shlens (2014), https://arxiv.org/pdf/1404.1100.pdf

Unit 01 | Intro

Idea



단순히 차원이 크니까 줄이자!! 이게아니고
관측된 차원이 아닌
실제 data의 latent space(숨겨진 진짜 dimension)를 찾아서 데이터를 더 잘 이해하고자 하는 노력!!

FIG. 1 A toy example. The position of a ball attached to an oscillating spring is recorded using three cameras A, B and C. The position of the ball tracked by each camera is depicted in each panel below.

"A Tutorial on PCA", J. Shlens (2014), https://arxiv.org/pdf/1404.1100.pdf

Unit 02 | Principal Component Analysis |

개념

$$X_1, X_2, X_3, X_4 \ldots, X_p$$

주성분 분석

$$Z_1, Z_2$$

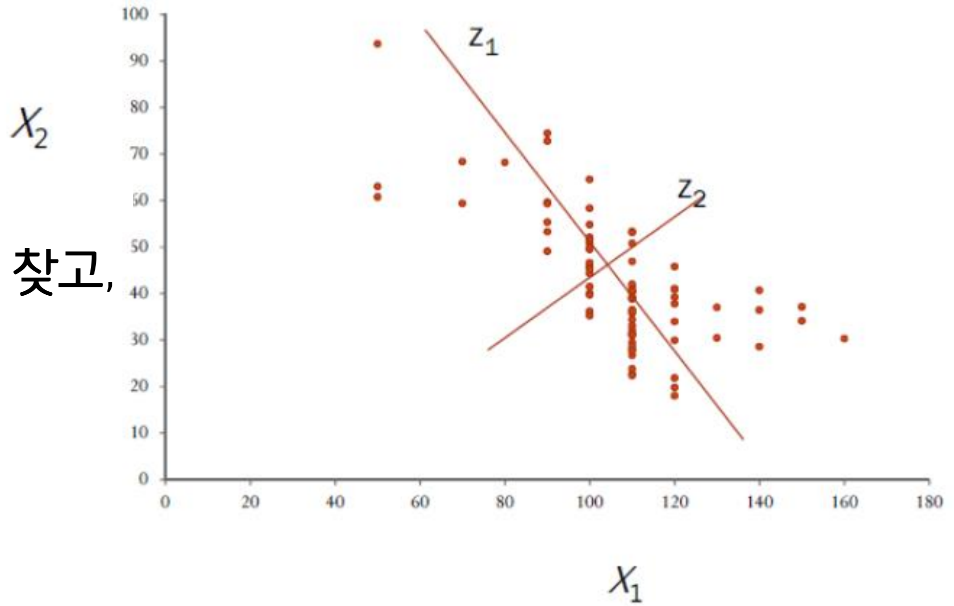
$$\vec{z_1} = \alpha_{11}\vec{x_1} + \alpha_{12}\vec{x_2} + \ldots + \alpha_{1p}\vec{x_p} = \vec{\alpha_1}^T X$$
$$\vec{z_2} = \alpha_{21}\vec{x_1} + \alpha_{22}\vec{x_2} + \ldots + \alpha_{2p}\vec{x_p} = \vec{\alpha_2}^T X$$
$$\ldots$$
$$\vec{z_p} = \alpha_{p1}\vec{x_1} + \alpha_{p2}\vec{x_2} + \ldots + \alpha_{pp}\vec{x_p} = \vec{\alpha_p}^T X$$

$$Z = \begin{bmatrix} \vec{z_1} \\ \vec{z_2} \\ \ldots \\ \vec{z_p} \end{bmatrix} = \begin{bmatrix} \vec{\alpha_1}^T X \\ \vec{\alpha_2}^T X \\ \ldots \\ \vec{\alpha_p}^T X \end{bmatrix} = \begin{bmatrix} \vec{\alpha_1}^T \\ \vec{\alpha_2}^T \\ \ldots \\ \vec{\alpha_p}^T \end{bmatrix} X = A^T X$$

Unit 02 | Principal Component Analysis |

그렇다면 어떤 dimension ?

Data의 Variance가 가장 큰 첫번째 축(첫 번째 주성분)을 찾고,
그 축에 직교하면서 또 가장 Var가 큰 그 다음 축(두 번째 주성분)을 찾고,
그 축에 직교하면서….하는 식으로 주성분을 생성한다.



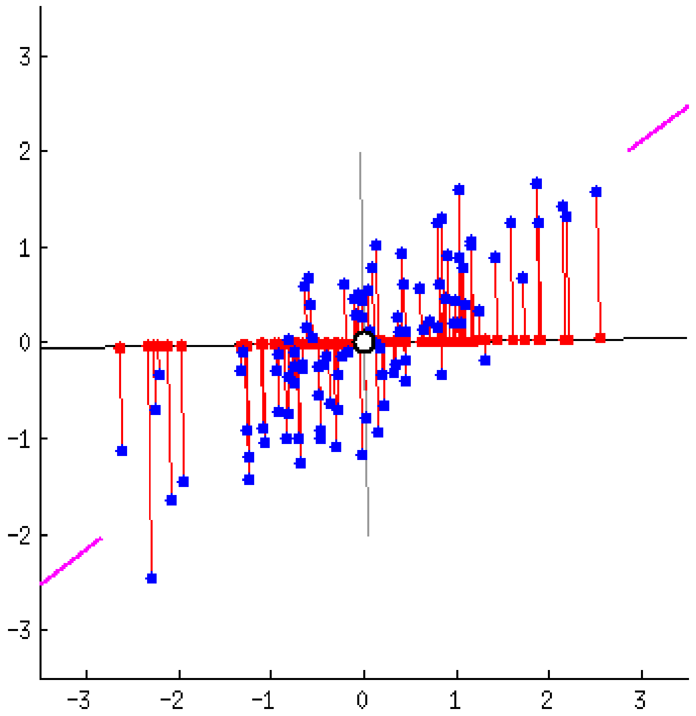
Why ?

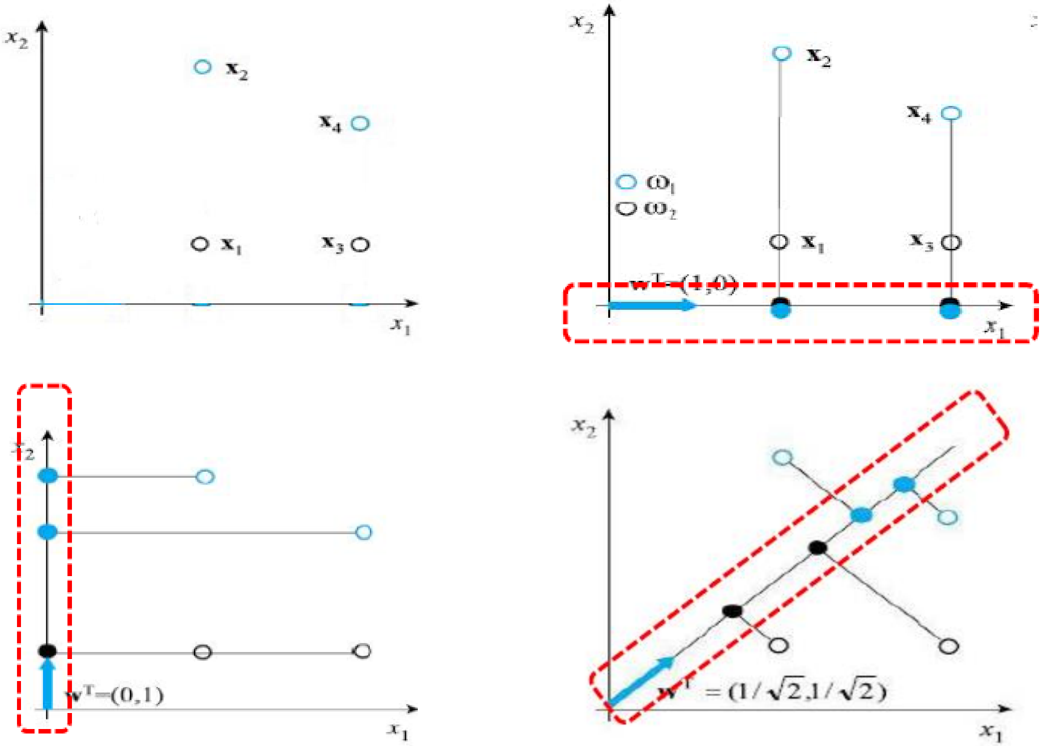
⇒ Z_1 : 전체 변수들을 가장 잘 설명해줄 수 있는 선형결합 벡터
Z_2 : 나머지 정보를 겹치는 부분 없이 가장 잘 설명해줄 수 있는 선형결합 벡터
⇒ 큰 분산을 갖는 방향이 중요한 정보를 담고 있다고 가정하기 때문 ! (왜 ?)
⇒ 분산이 가장 큰 방향(축) 위로 data들을 사영시키는 것은 선형결합과 같은 의미이다. (설명 ?)

https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Unit 02 | Principal Component Analysis |

왜 var가 큰 축이 data를 잘 설명해 줄 수 있는거지 ?





https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Unit 02 | Principal Component Analysis |

그 축은 어떻게 찾는거지?

Data ⇒ covariance matrix 생성 ⇒ eigen value & eigen vector 찾기 ⇒ 그 중 몇 개만 선택 !

그 축은 어떻게 찾는거지?

Data ⇒ covariance matrix 생성 ⇒ eigen value & eigen vector 찾기 ⇒ 그 중 몇 개만 선택 !

무엇이 eigen value & eigen vector인지 ?

왜 eigen value & eigen vector인지 ?

Unit 03 | Principal Component Analysis II

고유값과 고유벡터

Definition) Eigenvalues and Eigenvectors

Let A be an n x n matrix. A number λ is said to be an eigenvalue of A if there exists a nonzero solution vector K of the linear system

$$AK = \lambda K$$

The solution vector K is said to be an eigenvector corresponding to the eigenvalue λ.

Unit 03 | Principal Component Analysis II

고유값과 고유벡터

Definition) Eigenvalues and Eigenvectors

Let A be an n x n matrix. A number λ is said to be an eigenvalue of A if there exists a nonzero solution vector K of the linear system

$$AK = \lambda K$$

The solution vector K is said to be an eigenvector corresponding to the eigenvalue λ.

A : 정방행렬

λ : 행렬 A의 고유값 (Eigenvalue)

K : 행렬 A의 λ에 대한 고유벡터 (Eigenvector)

Nonzero solution ?

고유값과 고유벡터

Definition) Eigenvalues and Eigenvectors

Let A be an n x n matrix. A number λ is said to be an eigenvalue of A if there exists a nonzero solution vector K of the linear system

$$AK = \lambda K$$

The solution vector K is said to be an eigenvector corresponding to the eigenvalue λ.

$AK - \lambda K = 0$

$(A - \lambda I)K = 0 \Rightarrow (A - \lambda I)$가 역행렬을 가지면 K는 무조건 0

$\Rightarrow \det(A - \lambda I) = 0$이어야 함 !

고유값과 고유벡터 간단한 예제

$(A-\lambda I)K = 0 \Rightarrow (A-\lambda I)$가 역행렬을 가지면 K는 무조건 0

$\Rightarrow \det(A-\lambda I) = 0$이어야 함 !

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$|A-\lambda I| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 1 = 0$$

고유값과 고유벡터 간단한 예제

$(A - \lambda I)K = 0 \Rightarrow (A - \lambda I)$가 역행렬을 가지면 K는 무조건 0

$\Rightarrow \det(A - \lambda I) = 0$이어야 함 !

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 1 = 0$$

$$(2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3) = 0$$

Unit 03 | Principal Component Analysis II

고유값과 고유벡터 간단한 예제

$(A - \lambda I)K = 0 \Rightarrow (A - \lambda I)$가 역행렬을 가지면 K는 무조건 0

$\Rightarrow \det(A - \lambda I) = 0$이어야 함 !

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 1 = 0$$

$$(2-\lambda)^2 - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3) = 0$$

$\lambda_1 = 1,\ \lambda_2 = 3 \quad \Rightarrow$ A의 Eigenvalue는 1, 3 !!

Unit 03 | Principal Component Analysis II

고유값과 고유벡터 간단한 예제

$\lambda_1 = 1,\ \lambda_2 = 3$ ⇒ A의 Eigenvalue는 1, 3 !!

Definition) Eigenvalues and Eigenvectors

Let A be an n x n matrix. A number λ is said to be an eigenvalue of A if there exists a nonzero solution vector K of the linear system

$$AK = \lambda K$$

The solution vector K is said to be an eigenvector corresponding to the eigenvalue λ.

1) $\lambda_1 = 1$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

고유값과 고유벡터 간단한 예제

$\lambda_1 = 1,\ \lambda_2 = 3 \ \Rightarrow$ A의 Eigenvalue는 1, 3 !!

Definition) Eigenvalues and Eigenvectors

Let A be an n x n matrix. A number λ is said to be an eigenvalue of A if there exists a nonzero solution vector K of the linear system

$$AK = \lambda K$$

The solution vector K is said to be an eigenvector corresponding to the eigenvalue λ.

1) $\lambda_1 = 1$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

$$\begin{cases} 2v_1 + v_2 = v_1 \\ v_1 + 2v_2 = v_2 \end{cases} \Rightarrow \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

1) $\lambda_2 = 3$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

$$\begin{cases} 2v_1 + v_2 = 3v_1 \\ v_1 + 2v_2 = 3v_2 \end{cases} \Rightarrow \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

기하학적으로 이해하기

Definition) Eigenvalues and Eigenvectors

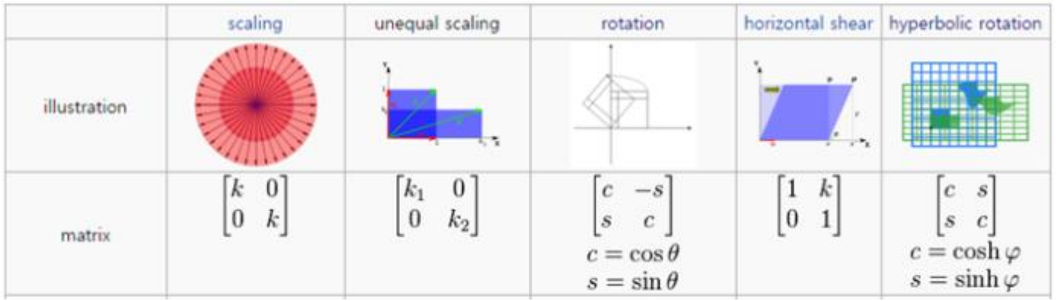
Let A be an n x n matrix. A number λ is said to be an eigenvalue of A if there exists a nonzero solution vector K of the linear system

$$AK = \lambda K$$

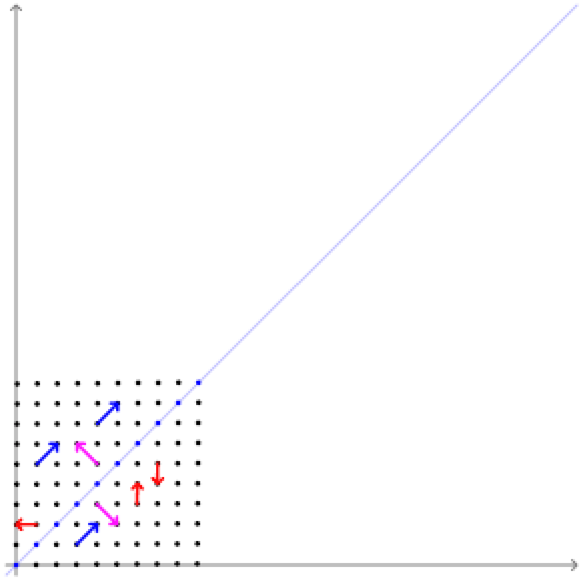
The solution vector K is said to be an eigenvector corresponding to the eigenvalue λ.

$$AK = \lambda K$$

$AK \Rightarrow$ 벡터에 선형변환



	scaling	unequal scaling	rotation	horizontal shear	hyperbolic rotation
illustration					
matrix	$\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}$	$\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$	$\begin{bmatrix} c & -s \\ s & c \end{bmatrix}$ $c = \cos\theta$ $s = \sin\theta$	$\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} c & s \\ s & c \end{bmatrix}$ $c = \cosh\varphi$ $s = \sinh\varphi$



Unit 03 | Principal Component Analysis II

기하학적으로 이해하기

Definition) Eigenvalues and Eigenvectors

Let A be an n x n matrix. A number λ is said to be an eigenvalue of A if there exists a nonzero solution vector K of the linear system

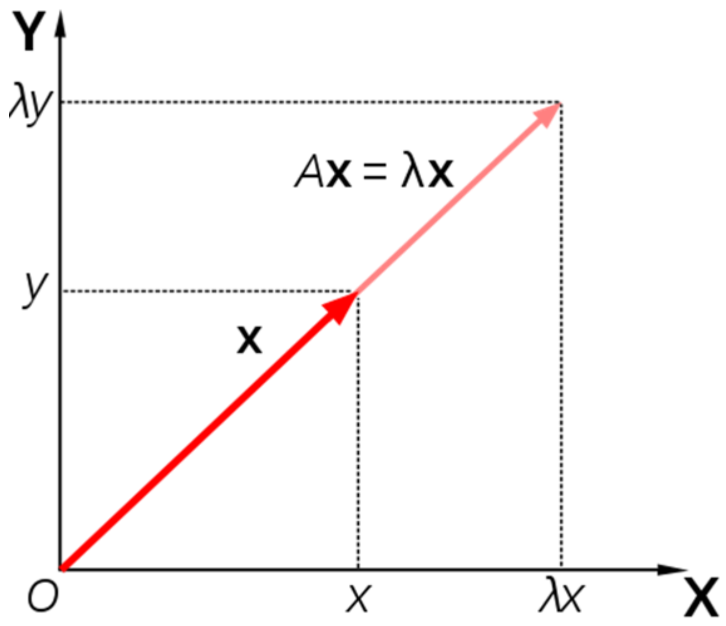
$$AK = \lambda K$$

The solution vector K is said to be an eigenvector corresponding to the eigenvalue λ.

$$AK = \lambda K$$

$AK \Rightarrow$ 벡터에 선형변환

$\lambda K \Rightarrow$ 벡터에 상수배



Unit 03 | Principal Component Analysis II

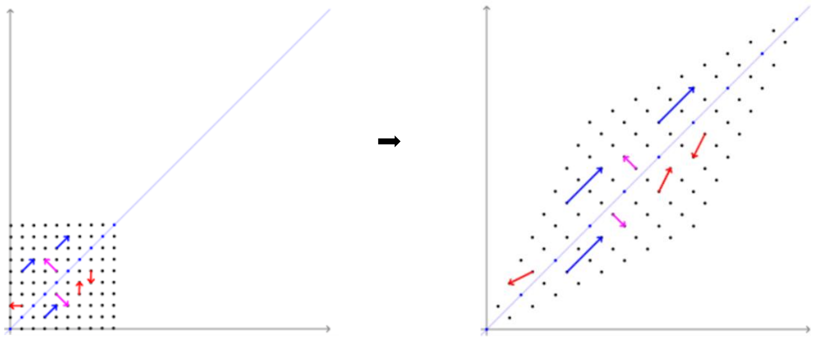
기하학적으로 이해하기

Definition) Eigenvalues and Eigenvectors

Let A be an n x n matrix. A number λ is said to be an eigenvalue of A if there exists a nonzero solution vector K of the linear system

$$AK = \lambda K$$

The solution vector K is said to be an eigenvector corresponding to the eigenvalue λ.



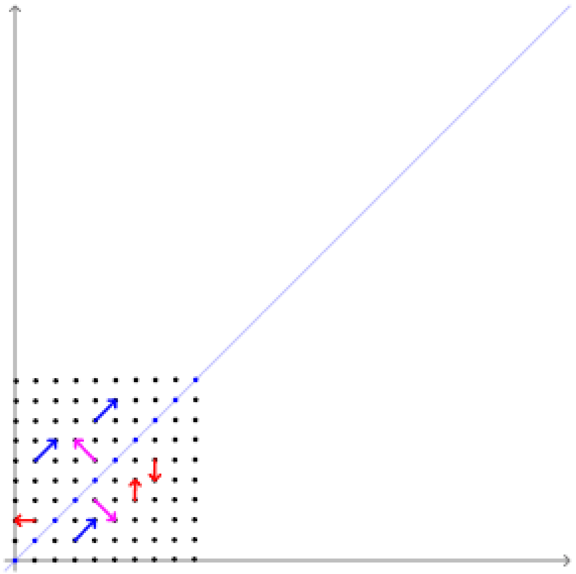
$$AK = \lambda K$$

$AK \Rightarrow$ 벡터에 선형변환

$\lambda K \Rightarrow$ 벡터에 상수배

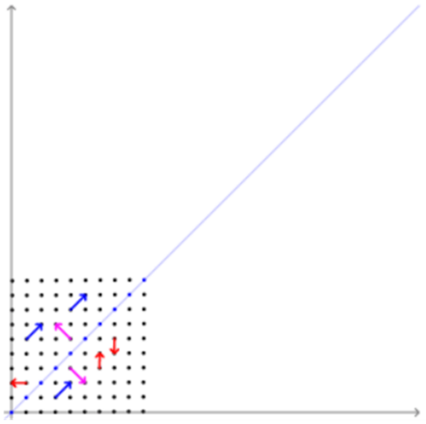
$\Rightarrow$ 어떠한 선형변환 A가 있을 때, 크기만 변하고 방향이 변하지 않는 벡터가 있나요?

그리고 그 벡터의 크기는 얼마만큼 변했나요?

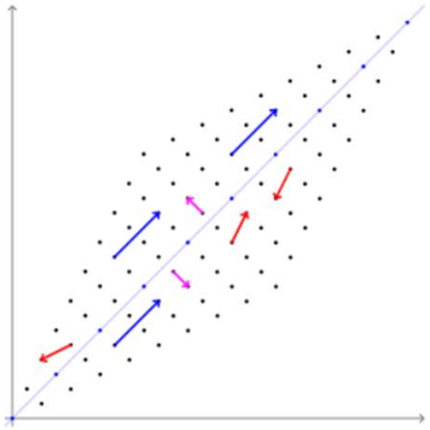


Unit 03 | Principal Component Analysis II

기하학적으로 이해하기



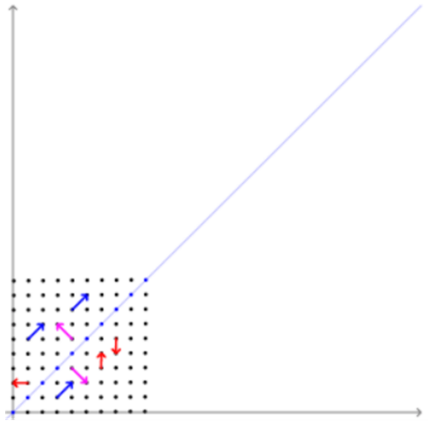
$$\rightarrow \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \rightarrow$$



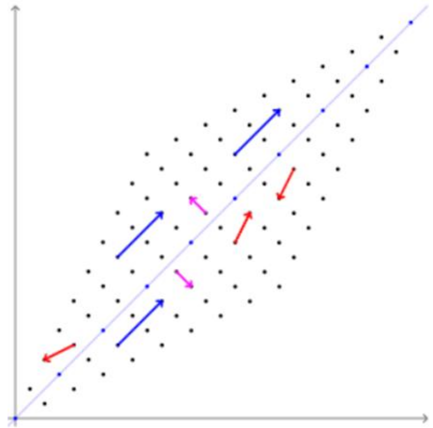
$\lambda_1 = 1$ 일 때 eigenvector $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$은 선형변환 A 취해주면 크기는 1배 되었지만 방향은 변하지 않는다.

$\lambda_2 = 3$ 일 때 eigenvector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$은 선형변환 A 취해주면 크기는 3배 되었지만 방향은 변하지 않는다.

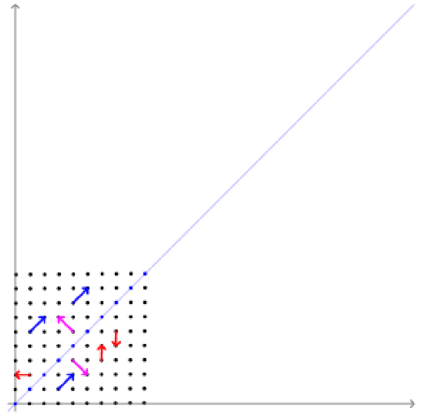
기하학적으로 이해하기



$\rightarrow \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \rightarrow$



" 어떤 변환 시 변환의 주축이 되는 축을 찾는 문제에 확장 시킬 수 있다 ! "



$\lambda_1 = 1$ 일 때 eigenvector $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$은 선형변환 A 취해주면 크기는 1배 되었지만 방향은 변하지 않는다.

$\lambda_2 = 3$ 일 때 eigenvector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$은 선형변환 A 취해주면 크기는 3배 되었지만 방향은 변하지 않는다.

빨간색 벡터들은 크기와 방향이 모두 바뀌기 때문에 Eigenvector가 아니다.

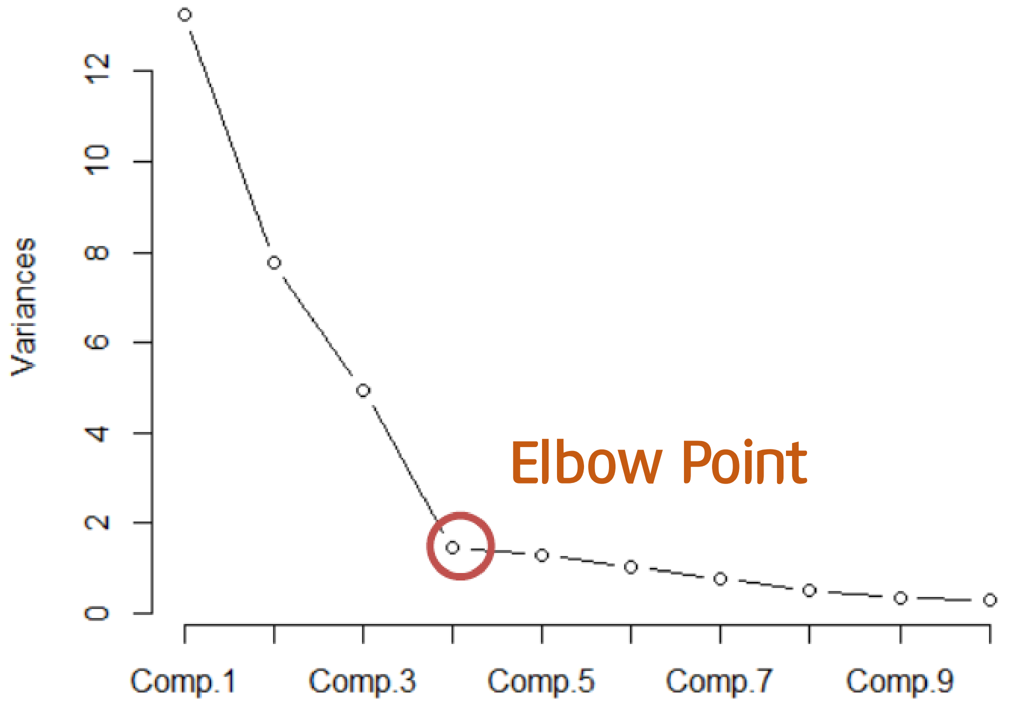
Unit 03 | Principal Component Analysis II

몇 개의 주성분을 사용하지 ?

Data ⇒ covariance matrix 생성 ⇒ Eigen value & Eigen vector 찾기 ⇒ 그 중 몇 개만 선택 !

Unit 03 | Principal Component Analysis II

몇 개의 주성분을 사용하지 ?



< 일반적 기준 >

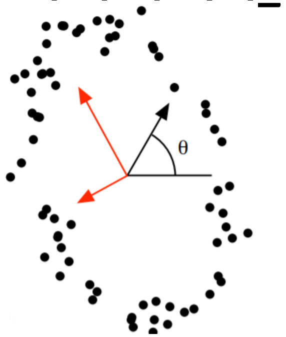
- Elbow Point 또는 이전 까지의 주성분을 사용
- Cumulative Proportion가 70~80% 이상

Unit 03 | Principal Component Analysis II

장점

1. 변수간 상관관계, 연관성을 이용해 차원을 축소 ⇒ data 이해 & 관리에 good !
2. 다중공선성의 해결 !
   ⇒ 변수들 사이에 강한 선형관계가 성립하는 경우 발생하는 다중 공선성을 주성분으로 축소
   ⇒ 변수제거 시 오는 정보 손실, 분석자의 주관적 판단 등으로 인한 위험 감소

한계

1. Linearity assumption (ex. )

2. Variance가 크다고 꼭 중요하다고 할 수 없다 (ex. Normalization의 중요성)
3. Orthogonal assumption (ex. 무엇이 더 중요한 축인가)

"A Tutorial on PCA", J. Shlens (2014), https://arxiv.org/pdf/1404.1100.pdf

Unit 02 | Principal Component Analysis

장점

1. 변수간 상관관계, 연관성을 이용해 차원을 축소 ⇒ data 이해 & 관리에 good !

한계에는 가정에 대해서만 적어 놓았는데,
이밖에도 새로 만든 주성분을 해석하기 어렵다는 단점이 있어요.
해석은 해당분야의 전문가와 상의하여 보는 것이 좋다고 합니다 !

1. Linearity assumption (ex. )
2. Variance가 크다고 꼭 중요하다고 할 수 없다 (ex. Normalization의 중요성)
3. Orthogonal assumption (ex. 무엇이 더 중요한 축인가)

"A Tutorial on PCA", J. Shlens (2014),
https://arxiv.org/pdf/1404.1100.pdf

Unit 03 | Principal Component Analysis II

Other Dimension Reduction Tech.

Dimension Reduction Technic의 기반인 PCA를 배웠다.
앞장의 한계라고 되어있는 부분은 PCA의 가정 !
Dimension reduction tech 마다 다른 가정을 하고 있기 때문에,
어떤 가정을 두고, 무엇을 줄이고, 어떤 정보를 최대화할지 또는 줄일지 잘 알고 data에 적용하는 것이 매우 중요하다.

대표적 차원축소법

t-SNE, Auto encoder, Kernel PCA 등

Unit 03 | Principal Component Analysis II

Other Dimension Reduction Tech.

Dimension Reduction Technic의 기반인 PCA를 배웠다.
앞장의 한계라고 되어있는 부분은 PCA의 가정 !
Dimension reduction tech 마다 다른 가정을 하고 있기 때문에,
어떤 가정을 두고, 무엇을 줄이고, 어떤 정보를 최대화할지 또는 줄일지 잘 알고 data에 적용하는 것이 매우 중요하다.

대표적 차원축소법

t-SNE, Auto encoder, Kernel PCA 등 ⇒ 장재석 !!!

실습 전에 !

1. Data 불러오기

Unit 04 | PCA in R

실습 전에 !

1. Data 불러오기
2. Scale : 각 변수에 대해 평균과 표준편차로 표준화된 data 만들기

Unit 04 | PCA in R

실습 전에 !

1. Data 불러오기
2. Scale : 각 변수에 대해 평균과 표준편차로 표준화된 data 만들기
3. Scaled data로 상관계수행렬 체크 : 다중공선성 여부, ※ Y 빼고 볼 것 !

Unit 04 | PCA in R

실습 전에 !

1. Data 불러오기
2. Scale : 각 변수에 대해 평균과 표준편차로 표준화된 data 만들기
3. Scaled data로 상관계수행렬 체크 : 다중공선성 여부, ※ Y 빼고 볼 것 !
4. prcomp 함수로 주성분 만들기 ※ princomp 함수 ?

Unit 04 | PCA in R

실습 전에 !

수행하는 알고리즘과 매개변수의 차이

Prcomp

- 원 자료에 대한 SVD(Singular Value Decomposition)을 이용
- 더 정확한 결과

Princomp

- 상관행렬에 대해 고유값분해 사용
- 더 많은 결과 출력
- 정량적 주성분 분석만 가능

	prcomp()	princomp()
square root of the eigenvalues	$sdev	$sdev
eigenvectors	$rotation	$loadings
means of each variable	$center	$center
scaling used of FALSE	$scale	$scale
principal components	$x	$scores
number of observation of each variable		$n.obs
the call that crated the object		$call

Unit 04 | PCA in R

실습 전에 !

1. Data 불러오기
2. Scale : 각 변수에 대해 평균과 표준편차로 표준화된 data 만들기
3. Scaled data로 상관계수행렬 체크 : 다중공선성 여부, ※ Y 빼고 볼 것 !
4. prcomp 함수로 주성분 만들기 ※ princomp 함수 ?
5. Elbow point로 몇 개의 주성분 사용할지 결정하기

Unit 04 | PCA in R

실습 전에 !

1. Data 불러오기
2. Scale : 각 변수에 대해 평균과 표준편차로 표준화된 data 만들기
3. Scaled data로 상관계수행렬 체크 : 다중공선성 여부, ※ Y 빼고 볼 것 !
4. prcomp 함수로 주성분 만들기 ※ princomp 함수 ?
5. Elbow point로 몇 개의 주성분 사용할지 결정하기
6. [주성분 회귀분석] 만들어진 주성분을 설명변수로 Y에 회귀시키기

Unit 04 | PCA in R

실습 전에 !

1. Data 불러오기
2. Scale : 각 변수에 대해 평균과 표준편차로 표준화된 data 만들기
3. Scaled data로 상관계수행렬 체크 : 다중공선성 여부, ※ Y 빼고 볼 것 !
4. prcomp 함수로 주성분 만들기 ※ princomp 함수 ?
5. Elbow point로 몇 개의 주성분 사용할지 결정하기
6. [주성분 회귀분석] 만들어진 주성분을 설명변수로 Y에 회귀시키기

⇒ 과제 : 오늘 배운 PCA와 2주차에 배웠던 회귀분석을 복습하는 의미에서 주성분 회귀분석을 해주세요~

실습 전에 !

1. Data 불러오기

끝으로…
PCA assumption이 맞다면, 최신 방법보다는 PCA가 제일 좋다고 함
가정 3가지에서 이렇게 이렇게 확인해 봐라 말씀드렸지만,
처음 data 받았을 때 한 번은 돌려보는 걸 추천 !

⇒ 과제 : 오늘 배운 PCA와 2주차에 배웠던 회귀분석을 복습하는 의미에서 주성분 회귀분석을 해주세요~

Q & A

들어주셔서 감사합니다.