머신러닝응용 제07강

Principal Component Analysis

첨단공학부 김동하교수

제07강 Principal Component Analysis

1	비지도 학습법 용어에 대해 학습한다.
2	차원축소기법 용어에 대해 학습한다.
3	주성분 변수의 유도 과정에 대해 학습한다.
4	주성분 개수를 결정하는 방법에 대해 학습한다.

핵심 단어

- 비지도 학습법
- 주성분 변수
- 고유값, 고유벡터
- Scree plot

07강. Principal Component Analysis

01. 비지도 학습법

1) 지도 학습법

- Supervised learning

- 사람이 교사로서 각각의 입력(X)에 대해 레이블(Y)를 달아놓은 데이터를 컴퓨터가 학습할 수 있도록 하는 방법

- 컴퓨터가 예측하는 것을 사람으로부터 교정받을 수 있음 -> 지도 학습

1) 지도 학습법

- 레이블의 형태에 따라
  - 연속형 변수 -> 회귀 (regression)
  - 이산형 변수 -> 분류 (classification)

2) 비지도 학습법

- Unsupervised learning
- 사람 없이 컴퓨터가 스스로 레이블이 없는 데이터에 대해서 학습.
- 즉, Y값 없이 X값만을 이용하여 학습.

2) 비지도 학습법

- 군집 분석 (Clustering analysis)
- 분포 추정 (Probability density estimation)
- 연관 분석 (Association analysis)

07강. Principal Component Analysis

02. 차원축소기법

1) 차원축소기법

- 고차원의 자료를 분석하기 위해서는 자료의 차원을 축소하는 것이 유리.

- 분석 자료의 주요 정보를 최대한 잃지 않으면서 변수의 수를 줄이는 방법
  - 차원축소기법

1) 차원축소기법

- **변수 선택 (Feature selection)**
  - 기존 변수 중 중요한 일부 변수만을 빼내는 기법.
- **변수 변환 (Feature transformation)**
  - 기존 변수를 조합해 새로운 변수를 만드는 기법.

2) 주성분분석

Principal Component Analysis (PCA)

- 대표적인 차원 축소 기법 중 하나
  - 변수 변환에 기초

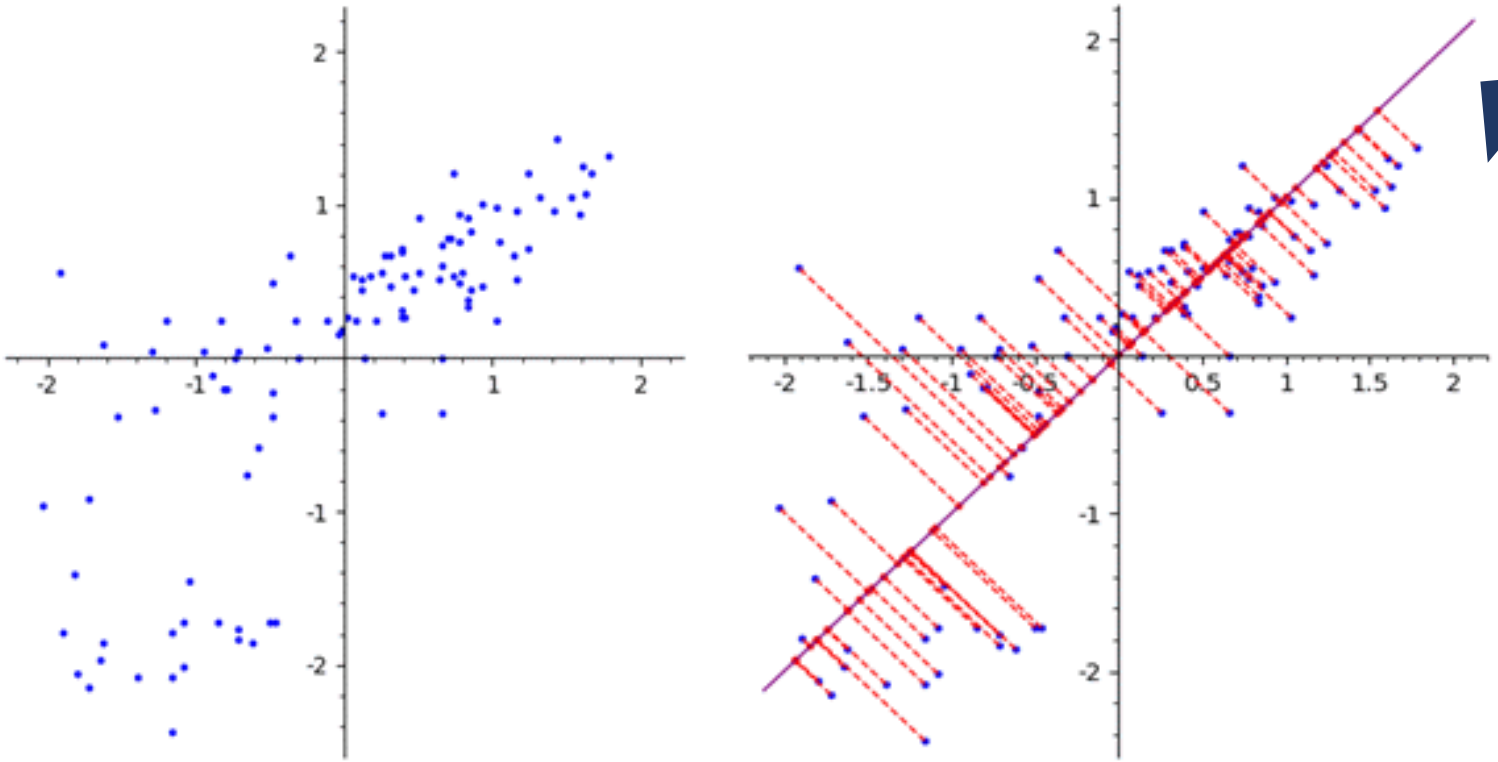
- 기존 변수들의 선형 변환을 통해 데이터를 잘 설명하는 새로운 변수들을 찾고, 이를 이용해 데이터의 차원을 축소.

2) 주성분분석

- Principal Component Analysis (PCA)
  - 원 데이터의 분산을 최대한 보존.
  - 서로 직교하는 주성분을 찾는 것이 핵심.

2) 주성분분석

Principal Component Analysis (PCA)



제 1주성분 (PC1)

출처: http://matrix.skku.ac.kr/math4ai-intro/W12/

07강. Principal Component Analysis

03. 주성분분석

1) 주성분 변수의 유도 과정

◆ 주성분 변수의 조건

- 원 데이터의 분산을 최대한 보존
- 주성분 변수끼리 직교

1) 주성분 변수의 유도 과정

제 1 주성분 (PC1)

- $X \in \mathbb{R}^{n \times p}$: p차원 자료 n개를 모은 행렬
  - 독립변수 데이터

- $x_i \in \mathbb{R}^p$: 자료 X의 i번째 자료 $(i = 1, \ldots, n)$

1) 주성분 변수의 유도 과정

제 1 주성분 (PC1)

- 벡터 $a \in \mathbb{R}^p$에 대해서 자료 x_i를 정사영(projection)했을 때의 좌표
  - $a^T x_i$

- n개의 자료를 모두 정사영했을 때 각각의 좌표
  - $a^T X \in \mathbb{R}^n$

1) 주성분 변수의 유도 과정

제 1 주성분 (PC1)

- 벡터 a가 자료 X의 분산을 잘 보존한다?
  - $a^T x_i, i = 1, \ldots, n$의 분산이 크다!

- $Var(a^T X)$를 최대화하는 벡터 a를 찾자!
  - 제 1 주성분 (PC1)
  - a의 크기를 1로 제한 (즉, $a^T a = 1$).

1) 주성분 변수의 유도 과정

◆제 1 주성분 (PC1)

- 라그랑지안 방법을 이용

$$v_1 = \underset{a}{argmax}\ \ a^T S a - \rho(a^T a - 1)$$

- S 는 자료 X 의 표본 공분산 행렬, ρ 는 라그랑지안 승수 (Lagrange multiplier).
  - $S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$
  - μ 는 자료 X 의 평균 벡터. ($\mu = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i$)

1) 주성분 변수의 유도 과정

◆제 1 주성분 v_1 은 표본 분산 행렬 S 의 가장 큰 고유값 (eigen value) 과 대응되는 고유 벡터 (eigen vector)가 됨.

- 고유값과 고유 벡터
  - 정방 행렬 A 의 고유값 (eigen value) 과 고유 벡터 (eigen vector) 는 다음 성질을 만족시키는 숫자 λ 와 벡터 v 를 의미.

$$A\mathbf{v} = \lambda\mathbf{v}$$

1) 주성분 변수의 유도 과정

◆제 2 주성분 (PC2)

◆제1 주성분과 직교하면서 $Var(a^T X)$ 를 최대화 하는 벡터 $a \in \mathbb{R}^p$ 를 찾는 것이 목표.

◆제 1 주성분과 마찬가지로 라그랑지안 방법을 사용.

$$\mathbf{v}_2 = \underset{a}{\operatorname{argmax}}\, a^T S a - \rho\left(a^T a - 1\right) - \phi a^T \mathbf{v}_1$$

- 여기서 ρ, ϕ 는 라그랑지안 승수.

1) 주성분 변수의 유도 과정

◆제 2 주성분 v_2 는 표본 분산 행렬 S 의 두 번째로 큰 고유값과 대응되는 고유 벡터임을 알 수 있음.

◆이와 같은 방법으로, 제 k 주성분은 표본 분산 행렬의 k번째로 큰 고유값과 대응되는 고유 벡터임을 보일 수 있음.

2) 차원 축소 데이터의 생성

- 자료 X 의 q 개의 주성분 $\mathbf{v}_1, \dots, \mathbf{v}_q$ 를 구함 $(q < p)$.
- $\mathbf{x}_i \in \mathbb{R}^p$: 자료 X 의 i 번째 자료. $(i = 1, \dots, n)$
- 주성분 분석을 통해 q 차원으로 차원 축소된 $\mathbf{x}_i$ 의 값은 $(\mathbf{x}_i^T \mathbf{v}_1, \dots, \mathbf{x}_i^T \mathbf{v}_q)$.

3) 주성분 개수 결정

최적의 주성분 개수 선정

- 주성분 중에서 데이터의 주요 정보를 갖고 있는 최적의 주성분 개수를 구해야 함.

- Scree Plot을 이용해 결정

3) 주성분 개수 결정

Scree Plot

- PCA 분석 결과를 이용해 고유벡터 방향의 분산 설명 정도를 나타낸 그림.

- 데이터 X의 주성분벡터를 $v_1, \dots, v_p$라 하면 j번째의 분산 설명 정도는 다음과 같이 계산되어짐.

$$\frac{Var(X^T v_j)}{\sum_{k=1}^{p} Var(X^T v_k)}$$

3) 주성분 개수 결정

Scree Plot

- 분산 변화율이 완만해지는 주성분의 수, 혹은 전체 분산의 70~90%가 되는 주성분의 수를 선정.

07강. Principal Component Analysis

04.Python을 이용한 실습

1) 데이터 설명

Fat dataset

- 252명 남성에 대한 나이, 몸무게, 키 등의 신체 정보와 비만도를 측정한 자료 (총 18가지의 변수).

2) 환경설정

◆필요한 패키지 불러오기

```
import os
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt

from sklearn import preprocessing
from sklearn.decomposition import PCA
```

3) 데이터 불러오기

◆필요한 패키지 불러오기

```
data_file = "./data/fat.csv"
fat = pd.read_csv(data_file)
print(fat.shape)
fat.head()
```

```
(252, 18)
```

	brozek	siri	density	age	weight	height	adipos	free
0	12.6	12.3	1.0708	23	154.25	67.75	23.7	134.9
1	6.9	6.1	1.0853	22	173.25	72.25	23.4	161.3
2	24.6	25.3	1.0414	22	154.00	66.25	24.7	116.0

4) 데이터 전처리

◆ 주성분분석을 위해서 데이터를 표준화한다.

```
fat_st = preprocessing.StandardScaler().fit_transform(fat)
feature_names = ['brozek', 'siri', 'density', 'age', 'weight',
                 'height', 'adipos', 'free', 'neck', 'chest',
                 'abdom', 'hip', 'thigh', 'knee', 'ankle', 'biceps',
                 'forearm', 'wrist']
fat_st = pd.DataFrame(fat_st, columns=feature_names)
fat_st.head()
```

4) 데이터 전처리

◆ 주성분분석을 위해서 데이터를 표준화한다.

	brozek	siri	density	age	weight
0	-0.819407	-0.820246	0.801647	-1.740073	-0.841246
1	-1.556273	-1.562573	1.565061	-1.819583	-0.193462
2	0.731890	0.736245	-0.746240	-1.819583	-0.849769
3	-1.039174	-1.047733	1.028039	-1.501543	0.198617
4	1.145569	1.143327	-1.135844	-1.660563	0.181570

5) 주성분분석

◆ 주성분분석을 시행한다.

```
pca = PCA(n_components = 18)
pca_components = pca.fit_transform(fat_st)
pca_fat = pd.DataFrame(data=pca_components,columns=
                       ['pc1','pc2','pc3','pc4','pc5','pc6',
                        'pc7','pc8','pc9','pc10','pc11','pc12',
                        'pc13','pc14','pc15','pc16','pc17','pc18'])
pca_fat.head()
```

5) 주성분분석

◆ 주성분분석을 시행한다.

	pc1	pc2	pc3	pc4	pc5
0	-2.686962	-0.588794	1.842965	0.423656	-0.245811
1	-1.821754	-3.120336	0.737673	0.263950	-0.103974
2	-1.891918	1.900563	2.911815	-0.893694	1.689662
3	-0.634725	-2.409151	0.825699	0.218584	-0.459951
4	0.904640	0.945138	2.304640	-1.769644	0.953840

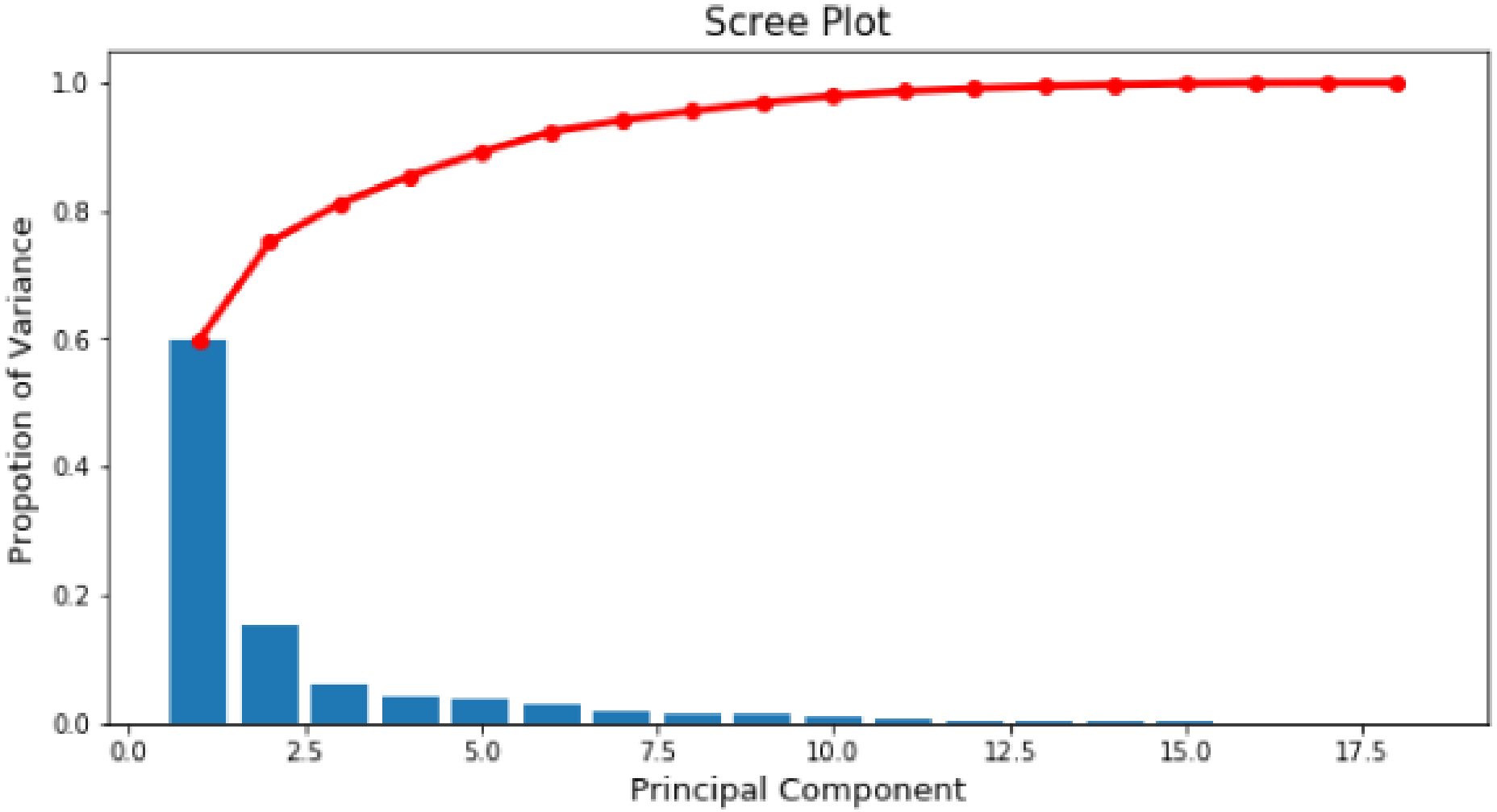
6) 최적의 주성분 수 계산

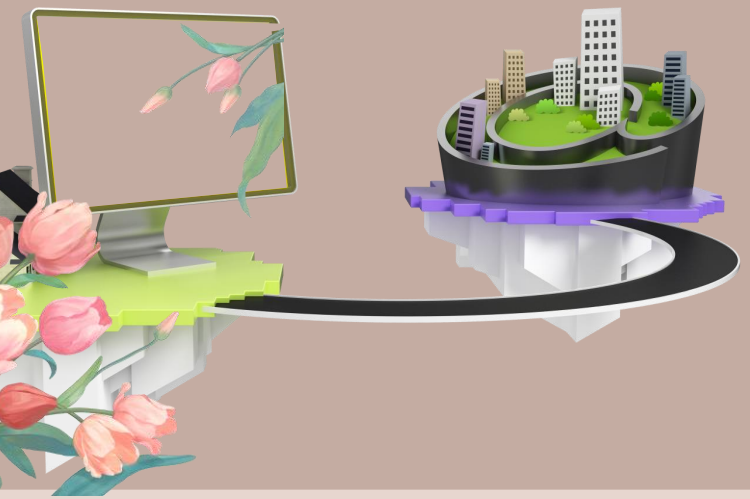
- Scree plot을 그려보자.

```
fig = plt.figure(figsize = (10, 5))
sing_vals = np.arange(18) + 1
vals = pca.explained_variance_ratio_
cumvals = np.cumsum(vals)
plt.bar(sing_vals, vals)
plt.plot(sing_vals, cumvals,
         'ro-', linewidth = 3)
plt.title('Scree Plot', fontsize=15)
plt.xlabel('Principal Component', fontsize=13)
plt.ylabel('Propotion of Variance', fontsize=13)
```

6) 최적의 주성분 수 계산

- 3개 정도가 적당해 보인다.





다음시간안내

제08강

Clustering Analysis