7강. Principal Component Analysis

◈ 담당교수 : 김 동 하

■ 학습개요

이번 강의에서는 자료의 선형변환을 통해 자료의 성질을 보존하면서 차원축소를 할 수 있는 방법론인 주성분분석에 대해서 배운다. 주성분과 고유벡터의 관계를 밝히고, 이를 통해 만들어진 차원축소된 자료의 형태를 파악한다. 더 나아가, 최적의 주성분 개수를 찾기 위한 Scree plot에 대해서도 학습한다.

■ 학습목표

1	비지도 학습법 용어에 대해 학습한다.
2	차원축소기법 용어에 대해 학습한다.
3	주성분 변수의 유도 과정에 대해 학습한다.
4	Scree Plot에 대해 학습한다.

■ 주요용어

용어	해설
비지도 학습법	지도 학습법과는 다르게 컴퓨터가 스스로 라벨이 없는 데이터에 대해서 학습하는 것을 총칭한다. 비지도 학습법의 예로는 군집 분석, 연관 분석 등이 있다.
주성분 변수	주성분분석에서 새롭게 사용하는 방향벡터로 데이터의 분산을 최대한 보존하면서 타 주성분 변수들과는 직교하는 방향으로 설정한다.
고유값, 고유벡터	선형 변환이 일어난 후에도 방향이 변하지 않는 0이 아닌 벡터를 고유벡터라 하며, 고유 벡터의 길이가 변하는 배수를 고유값이라 한다.
Scree plot	주성분분석 후 주성분의 수를 선정하기 위해 고유값-주성분의 분산 변환 변화를 보는 그래프.

■ 학습하기

01. 비지도 학습법

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

비지도 학습법 (Unsupervised learning)

- 사람 없이 컴퓨터가 스스로 레이블이 없는 데이터에 대해서 학습.
- 즉, Y값 없이 X값만을 이용하여 학습.
- 비지도 학습의 대표적인 분석
  -> 군집 분석 (Clustering analysis)
  -> 분포 추정 (Probability density estimation)
  -> 연관 분석 (Association analysis)

02. 차원 축소 기법

차원 축소 기법

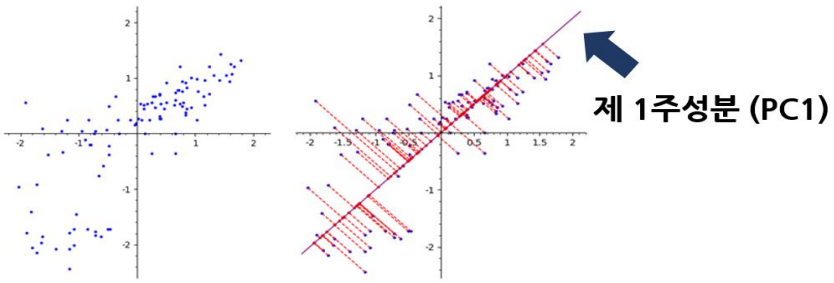
- 고차원의 자료를 분석하기 위해서는 자료의 차원을 축소하는 것이 유리.
- 분석 자료의 주요 정보를 최대한 잃지 않으면서 변수의 수를 줄이는 방법
  -> 차원 축소 기법

차원 축소 기법의 두 가지 접근

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

주성분 분석

- 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.
- 원 데이터의 분산을 최대한 보존.
- 서로 직교하는 주성분을 찾는 것이 핵심.



03. 주성분 분석

주성분 변수의 유도 과정

- 주성분 변수의 조건
  -> 1) 원 데이터의 분산을 최대한 보존하면서,
  -> 2) 주성분 변수끼리 직교해야 함.
- $X \in R^{n \times n}$: p차원 자료 n개를 모은 행렬
  -> 독립변수 데이터
- $x_i \in R^p$: 자료 X의 i번째 자료 $(i = 1,...,n)$

제 1 주성분 찾기 (1)

- 벡터 $a \in R^p$에 대해서 자료 x_i를 정사영 (projection)했을 때의 좌표는 $a^T x_i$.
- n개의 자료를 모두 정사영 했을 때의 각각의 좌표를 벡터 형태로 표시하면 다음과 같음:
  -> $a^T X \in R^n$
- 벡터 a가 자료 X의 분산을 잘 보존한다는 것은?
  -> $a^T x_i$, $i = 1,...,n$의 분산이 크다는 것.
- $Var(a^T X)$를 최대화하는 벡터 a를 찾자!
  -> 제 1 주성분 (PC1)
  -> a의 크기를 1로 제한 (즉, $a^T a = 1$)

제 1 주성분 찾기 (2)

- 라그랑지안 방법을 이용
  -> $v_1 = \arg max_a a^T S a - \rho(a^T a - 1)$
- 여기서, S는 자료 X의 표본 공분산 행렬, ρ는 라그랑지안 승수 (Lagrange multimplier).
  -> $S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$
  -> μ는 자료 X의 평균 벡터 (즉, $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$).

- 즉, 제 1 주성분 v_1은 표본 분산 행렬 S의 가장 큰 고유값 (eigen value)에 대응되는 고유 벡터 (eigen vector)가 됨.
- 고유값과 고유 벡터
  -> 정방 행렬 A의 고유값 (eigen value)과 고유 벡터 (eigen vector)는 다음 성질을 만족시키는 숫자 λ와 벡터 v를 의미.
  $Av = \lambda v$

제 2 주성분 찾기

- 제 1 주성분과 직교하면서 $Var(a^T X)$를 최대화하는 벡터 $a \in R^p$를 찾는 것이 목표.
- 제 1 주성분과 마찬가지로 라그랑지안 방법을 사용.
  -> $v_2 = \arg max_a a^T S a - \rho(a^T a - 1) - \phi a^T v_1$
- 여기서 ρ, ϕ는 라그장지안 승수.
- 제 2 주성분 v_2는 표본 분산 행렬 S의 두 번째로 큰 고유값과 대응되는 고유 벡터임이 알려져 있음.
- 이와 같은 방법으로, 제 k 주성분은 표본 분산 행렬의 k번째로 큰 고유값과 대응되는 고유 벡터임을 보일 수 있음.

차원 축소 데이터의 생성

- 자료를 q차원으로 축소하고 싶다고 가정하자 ($q < p$).
- 자료 X의 q개의 주성분 v_1, ..., v_q를 구함.
- $x_i \in R^p$: 자료 X의 i번째 자료. ($i = 1,...,n$)
- 주성분 분석을 통해 q차원으로 차원 축소된 x_i의 값은 다음과 같음:
  $\left(x_i^T v_1, ..., x_i^T v_q\right)$

주성분 개수의 결정

- 최적의 주성분 개수 선정
- 주성분 중에서 데이터의 주요 정보를 갖고 있는 최적의 주성분 개수를 구해야 함.
- Scree Plot을 이용해 결정

Scree plot

- PCA 분석 결과를 이용해 고유벡터 방향의 분산 설명 정도를 나타낸 그림.
- 데이터 X의 주성분벡터를 $v_1,...,v_p$라 하면 j번째의 분산 설명 정도는 다음과 같이 계산할 수 있음:
  $$\frac{Var\left(X^T v_j\right)}{\sum_{k=1}^{p} Var\left(X^T v_k\right)}$$
- 분산 변화율이 완만해지는 주성분의 수, 혹은 전체 분산의 70~90%가 되는 주성분의 수

를 선정.

04. Python을 이용한 실습

데이터 설명 (Fat data set)

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

환경 설정

- 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
```

데이터 불러오기 및 전처리

```
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

- 주성분분석을 위해 데이터를 표준화하는 작업이 선행되어야 한다.

```
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()
```

주성분분석

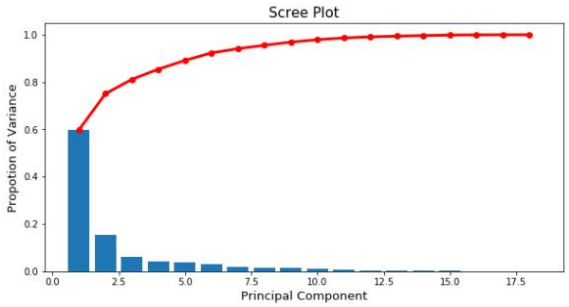
- 주성분분석을 시행한다.

```
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()
```

- Scree plot을 그려서 최적의 주성분 수를 계산해보자.

```
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)
```

- 3개가 적당해보인다.



■ 연습문제

(객관식)1. 비지도 학습법에 속하는 문제가 아닌 것을 고르시오.

① 군집 분석

② 분포 추정

③ 회귀 문제

④ 연관 분석

정답 : ③

해설 : 회귀 문제는 지도 학습법에 속하는 문제이다.

(O/X)2. 제 3 주성분은 제 1, 2 주성분과 직교하면서 데이터의 분산을 최대로 보존하는 방향이다.

정답) O

해설) 제 3 주성분은 제 1,2 주성분과 모두 직교해야 한다.

(단답형)3. 최적의 주성분 개수를 설정하기 위해 살펴보는 그림으로 PCA 분석 결과를 이용해 고유벡터 방향의 분산 설명 정도를 나타낸 그림을 무엇이라 하는가?

정답 : Scree plot

해설 : 최적의 주성분 개수를 알아보기 위해서 Scree plot을 살펴본다.

■ 정리하기

1. 기계 학습에는 크게 지도 학습법과 비지도 학습법이 있다.

2. 주성분은 데이터의 분산을 최대화하는 방향으로 구해지며, 이는 데이터의 표본 분산 행렬의 고유벡터가 된다.

3. 최적의 주성분 개수를 결정하기 위해서 scree plot을 살펴본다.

■ 참고자료 (참고도서, 참고논문, 참고사이트 등)

박창이, 김용대, 김진석, 송종우, 최호식. 『R을 이용한 데이터마이닝』. 서울:교우사, 2018.