科学技術計算

第6回レポート

30114150　KIMHYUNWOO

01.24

課題1

本kadai1.pyプログラムでは資料として与えられた行列 $A = \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$ とその固有値のベクトル $eigenvalvector = [-1 \quad 5]$ 及び固有ベクトル行列 $eigenvecmatrix = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}$ を用いて、固有ベクトルおよび固有値が満たす以下の３つの性質に対する検証を行う。

1. $\boldsymbol{Av = \lambda v}$

2. 固有値の和は $\boldsymbol{A}$ のトレースに等しい

3. 固有値の積は $\boldsymbol{A}$ の行列式に等しい

まず、項目１ $\boldsymbol{Av = \lambda v}$ を検証するために行列Aと固有ベクトルの積と固有ベクトルと固有値のdot productの結果を比較する。その結果は以下の通りであり、同じ値であることが確認できる。

```
A @ eigenvecmatrix =
 [[ 2  5]
 [-1  5]]
eigenvecmatrix * eigenvalvector =
 [[ 2  5]
 [-1  5]]
```

続いて、固有値の和は $\boldsymbol{A}$ のトレースに等しいことを検証するためにnp.trace(A)関数で求めたTrace_Aの値と固有値をnp.sum(eigenvalvector)関数で足した値を比較した結果は以下の通りであり、その値が同じであることが確認できる。

```
Trace_A =  4
sum =  4
Trace of A matrix and sum of eigenvalue are same
```

最後に、固有値の積は $\boldsymbol{A}$ の行列式に等しいことを検証するためにnp.linalg.det(A)関数で求めた行列式の値と、np.prod(eigenvalvector)関数で求めた固有値の積を比較した結果は以下の通りであり、その結果が同じであることから固有ベクトルや固有値が３つの性質を満たしていることが分かる。検証できる。

```
Det_A =  -5.000000000000001
prodvalue =  -5
Determinant of A matrix and product of eigenvalue are same
```

課題2

本kadai2.pyプログラムでは固有値・固有ベクトルを求める関数numpy.linalg.eig()の機能を確認する。検証対象は課題１と同じ行列A及び固有ベクトル、固有値である。

まず、numpy.linalg.eig(A)で固有ベクトルvや固有値wを求める。固有ベクトルvは行列の形で求められるが値が整数ではなく実数の形で与えられるため、各列ごとに最小値で割って整数の比の行列で変換する作業を行う。この整数の比に変換する作業は行列中に０が入っていない場合のみ行う。

```
w,v = np.linalg.eig(A)
if(np.min(abs(v)) != 0):
v = v / np.min(abs(v), axis = 0)
```

固有値は整数の形で与えられるため別途の操作は行わず、最後に numpy.linalg.eig(A)で求めた値と既知の固有値及び固有ベクトル行列を比較する。その結果は以下の通りで、両値が一致することが分かる。

```
w :  [-1.  5.]
eigenvalvector :  [-1  5]
v :
 [[-2. -1.]
 [ 1. -1.]]
eigenvecmatrix :
 [[-2  1]
 [ 1  1]]
```

課題3

本pca.pyプログラムでは、講義資料３９ページの入力データに対して主成分分析を行い、その結果及び寄与率、累積寄与率などを主成分順位ことに求めて、その結果が講義資料の結果と一致しているかを確認する。

主成分分析を行うために配列データを引数として受けて寄与率及び各主成分の値を返すPCA関数を定義する。手続きは講義資料の手順に従って、分散共分散行列を求めてからその行列の固有値及び固有ベクトルの導出、最後にデータと固有ベクトルの行列積で最終的な主成分を求める仕組みである。

まず、データを備えてからPCA関数に渡す。 以後、後の行列操作のためにデータの形やデータの平均を求めてから、分散共分散行列のための行列Sを準備する。

```
surveydata = np.array([[8,9,4],[2,5,7],[8,5,6],[3,5,4],[7,4,9],[4,3,4],[3,6,8],[6,8,2],[5,4,5],[6,7,6]])
N, d = np.shape(data)
dimmean = np.sum(data,axis=0)/N
S = np.zeros((d,d))
```

続いて、分散共分散行列の各要素を求めてからeig関数で分散共分散行列の固有値及び固有ベクトルを求める。両方小数点以下の桁数は最大４桁に制限する。

```
for i in range(d):
        for j in range(d):
            S[i,j] = np.sum(np.dot((data[:,i]-dimmean[i]),(data[:,j]-dimmean[j])))/N
    eigV, eigM = np.linalg.eig(S)
    np.round(eigV,4,eigV)
    np.round(eigM,4,eigM)
```

最後に得られた固有値で各成分の寄与率を求めて、データに固有ベクトルを行列積することで主成分分析を終了する。

```
proportion = eigV/sum(eigV)
Z = (data-dimmean) @ eigM
```

以上の過程の結果は続く２つのページ通りであり、結果を見ると講義資料とは符号が異なっていることが分かる。これはeig関数の影響であるが固有ベクトルの符号も反転されているため、実質的には符号さえ反転させれば意味は通じると考えられる。

1st component equation :

-4.484700 = -0.5986*(8-5.2)-0.6830*(9-5.6)-0.4186*(4-5.5)

+3.088100 = -0.5986*(2-5.2)-0.6830*(5-5.6)-0.4186*(7-5.5)

-1.051500 = -0.5986*(8-5.2)-0.6830*(5-5.6)-0.4186*(6-5.5)

+0.845500 = -0.5986*(3-5.2)-0.6830*(5-5.6)-0.4186*(4-5.5)

+1.775400 = -0.5986*(7-5.2)-0.6830*(4-5.6)-0.4186*(9-5.5)

+1.415500 = -0.5986*(4-5.2)-0.6830*(3-5.6)-0.4186*(4-5.5)

+2.453200 = -0.5986*(3-5.2)-0.6830*(6-5.6)-0.4186*(8-5.5)

-3.799200 = -0.5986*(6-5.2)-0.6830*(8-5.6)-0.4186*(2-5.5)

+0.780600 = -0.5986*(5-5.2)-0.6830*(4-5.6)-0.4186*(5-5.5)

-1.022900 = -0.5986*(6-5.2)-0.6830*(7-5.6)-0.4186*(6-5.5)

2nd component equation :

-0.766630 = -0.5843*(8-5.2)+0.0148*(9-5.6)+0.8114*(4-5.5)

+1.081270 = -0.5843*(2-5.2)+0.0148*(5-5.6)+0.8114*(7-5.5)

-2.286430 = -0.5843*(8-5.2)+0.0148*(5-5.6)+0.8114*(6-5.5)

+2.589170 = -0.5843*(3-5.2)+0.0148*(5-5.6)+0.8114*(4-5.5)

-3.809130 = -0.5843*(7-5.2)+0.0148*(4-5.6)+0.8114*(9-5.5)

+1.876570 = -0.5843*(4-5.2)+0.0148*(3-5.6)+0.8114*(4-5.5)

-0.317230 = -0.5843*(3-5.2)+0.0148*(6-5.6)+0.8114*(8-5.5)

+2.045170 = -0.5843*(6-5.2)+0.0148*(8-5.6)+0.8114*(2-5.5)

+0.478070 = -0.5843*(5-5.2)+0.0148*(4-5.6)+0.8114*(5-5.5)

-0.890830 = -0.5843*(6-5.2)+0.0148*(7-5.6)+0.8114*(6-5.5)

3rd component equation :

+0.974830 = +0.5480*(8-5.2)-0.7303*(9-5.6)+0.4079*(4-5.5)

+1.464530 = +0.5480*(2-5.2)-0.7303*(5-5.6)+0.4079*(7-5.5)

-1.454970 = +0.5480*(8-5.2)-0.7303*(5-5.6)+0.4079*(6-5.5)

-0.177770 = +0.5480*(3-5.2)-0.7303*(5-5.6)+0.4079*(4-5.5)

-0.624070 = +0.5480*(7-5.2)-0.7303*(4-5.6)+0.4079*(9-5.5)

-2.219170 = +0.5480*(4-5.2)-0.7303*(3-5.6)+0.4079*(4-5.5)

+2.265230 = +0.5480*(3-5.2)-0.7303*(6-5.6)+0.4079*(8-5.5)

+0.184830 = +0.5480*(6-5.2)-0.7303*(8-5.6)+0.4079*(2-5.5)

-1.418470 = +0.5480*(5-5.2)-0.7303*(4-5.6)+0.4079*(5-5.5)

+1.005030 = +0.5480*(6-5.2)-0.7303*(7-5.6)+0.4079*(6-5.5)

```
[[-4.4847  -0.76663  0.97483]
 [ 3.0881   1.08127  1.46453]
 [-1.0515  -2.28643 -1.45497]
 [ 0.8455   2.58917 -0.17777]
 [ 1.7754  -3.80913 -0.62407]
 [ 1.4155   1.87657 -2.21917]
 [ 2.4532  -0.31723  2.26523]
 [-3.7992   2.04517  0.18483]
 [ 0.7806   0.47807 -1.41847]
 [-1.0229  -0.89083  1.00503]]

1st proportion :  0.512934497816594
1st cumulative proportion :  0.512934497816594
2nd proportion :  0.3233449781659389
2nd cumulative proportion :  0.8362794759825328
3rd proportion :  0.16372052401746726
3rd cumulative proportion :  1.0
```

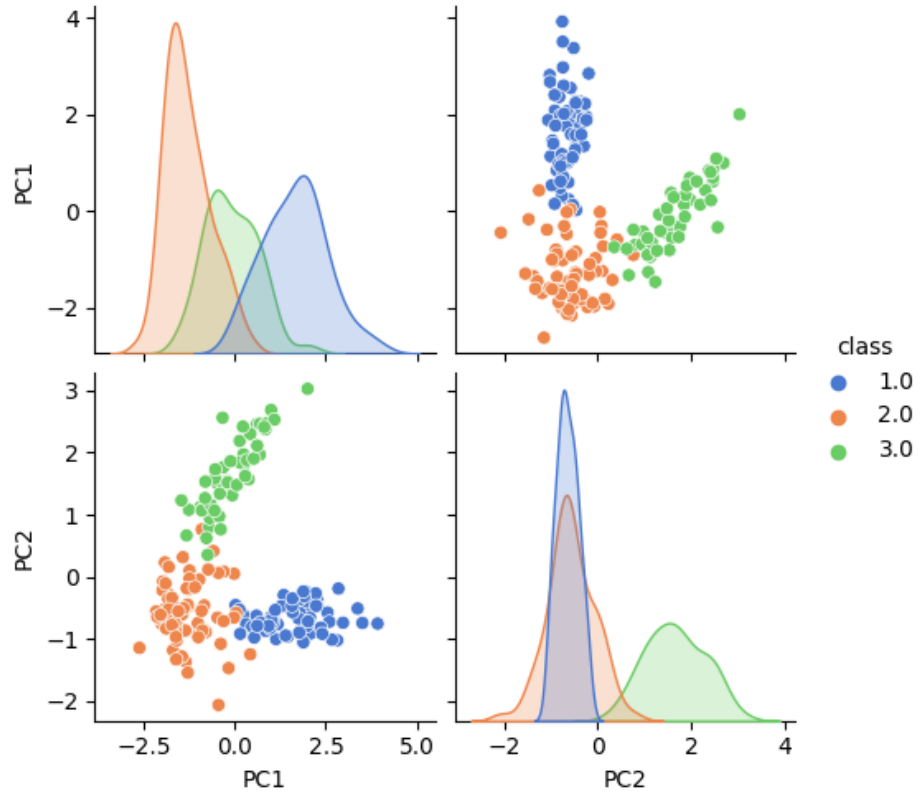
課題4

本kadai4.pyプログラムでは以下の５つの質問に対する答えを求める。

１）散布図行列から、４つの属性値のそれぞれがブドウの品種の違いをどの程度表せているか、概要を150字程度で述べよ。

散布図行列の対角成分であるヒストグラムから分析すると、１クラスがalcohol,flavanoids,prolin属性で最も高い数値を表しつつcolor　intensityも中間であることが分かる。２クラスはflavanoidsを除いた項目で最も低い数値を表しており、３クラスは　color　intensityのみ高い数値を見せることを除くと大体中間の数値を表すことが分かる。

２）主成分分析で第１主成分と第２主成分を求めプロットせよ。

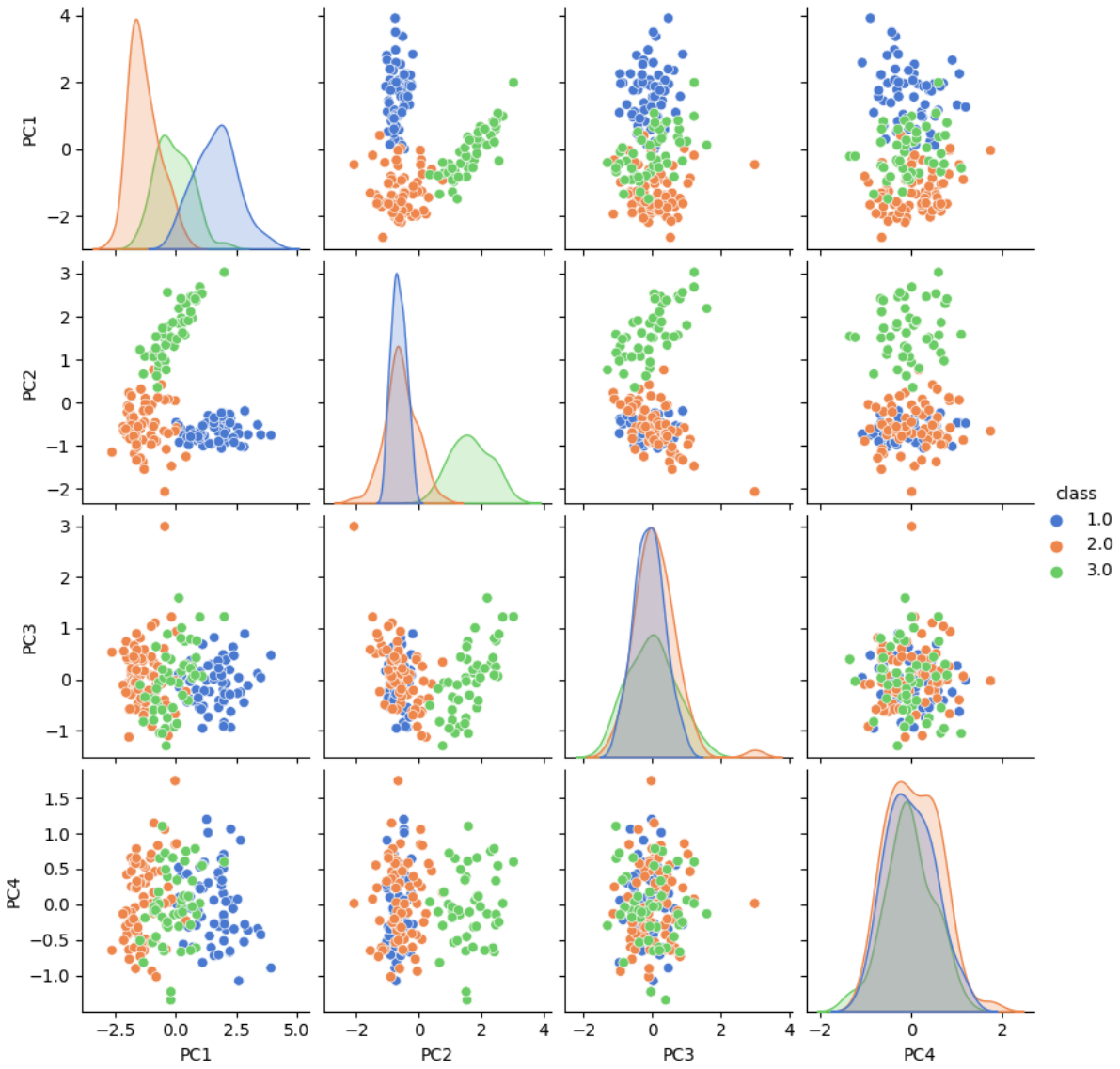


結果の図から各成分分析に対して射影したと考えると、第１主成分分析では２クラスが劣位で１クラスが優勢、３クラスは中間に位置していることが分かる。また、第２主成分分析では１，２クラスが差が大きくない反面、３クラスが優位を占めていることが分かる。これらの結果を項目１の品種別違いに関連付けて考察すると、第１主成分分析ではalcohol,flavanoids,proline成分に注目していて、第２主成分分析ではcolor intensityに注目していると考えられる。

３）第4主成分まで求め、各主成分（第1主成分～第4主成分）に対する散布図行列を出力せよ。

pca.fit_transform関数で求めた主成分分析の結果を講義資料で示した以下の関数と同様にプロットさせた結果は以下の通りである。

```
sns.pairplot(PCA_components, hue = "class", palette='muted', vars = ["PC1",'PC2','
PC3','PC4'])
```



４） 3）の散布図行列のうち「第1主成分 - 第2主成分ペア」と「第3主成分 – 第4主成分ペア」の図を比較せよ。そしてその差がどこからくるのか、PCA で得られる主成分の特性から論ぜよ。

```
[2.13412644 1.23807891 0.33914703 0.28864762]
[[ 0.60560099  0.32554588  0.40941882  0.59970288]
 [ 0.16828594 -0.7253591   0.63392917 -0.20896849]
 [-0.38871061  0.56689535  0.63654637 -0.34977484]
 [ 0.67367006  0.21564721 -0.15911867 -0.68872794]]
[0.53353161 0.30951973 0.08478676 0.07216191]
```

第１成分と第２成分ペアを比較して見ると分析結果が各クラス別に明確に区分されているが、第３成分と第４成分ペアではあまり区分が付けていないことが分かる。この様子を項目２の答えに関連付けて考察して見ると、固有ベクトル行列（パラメータ）の結果から実際に第１主成分分析ではalcohol,flavanoids,proline成分に注目していて、第２主成分分析ではcolor intensityに注目していることが分かる。一方、第３、４成分分析ではalcohol,flavanoids,color intensityかalcocholだけに注目していることが分かる。

以上の結果をまとめるとヒストグラムから把握した品種別違いに基づく項目１の分析通り実際にalcohol,flavanoids,proline成分かalcohol,color intensity,proline成分に注目した時、データの区分が明確になると考えられる。

５） 各主成分の寄与率を求めよ。

寄与率を計算するSklearnライブラリのpca.explained_variance_ratio_関数で寄与率を求めた結果は以下の通りである。

```
[0.53353161 0.30951973 0.08478676 0.07216191]
```

考察と感想

今回の主成分分析でデータから適切な特徴だけ選んでデータの次元を縮小させる一方、新しい基準から見たデータの分布分析仕方が分かった。データ間の分散共分散行列を求めてその分散が最大となる固有ベクトルを求める意味を考えると、分散が大きいほどその基準（軸）に射影したデータの分布が一番広く散らされている時、元データ別の構造を保ちながら次元が縮小できることと考えられる。

主成分分析に置いてその軸を選ぶ基準である固有値及び寄与率も重要であり、課題４で確認した通り余り関係ない主成分で射影するとデータ間の分布が混ざってしまうことから大きい固有値が意味する第ｎ主成分を選ぶ基準の意味も分かった。

主成分分析は画像の圧縮、特徴の抽出などの多方面で愛用される技法であるため、忘れてはいけない重要な技術であると考えらえる。