

統計分析法 第二回レポート

実行環境

```
python3.9
• pipenv
(これ以降のセクションは(pipenv update)でダウンロードできます)
• pandas
• matplotlib
• seaborn
```

ファイルのエンコーディングについて

utf8(uf8)、utf16(utf16), utf32(utf32), utf16le(utf16le)については、shf.pyが読み込みできそうですが、utf32(utf32), utf16le(utf16le)については、shf.pyでエンコーディングしてもエラーが返るため、utf8でエンコーディングしている。

主成分分析の流れについて

主成分分析はsklearn.decomposition.PCAを用いて行なった。

データについては、説明変数を5次元とした後、標準化して行っている。

また、図表、図表をそれぞれ、sklearn.decomposition.PCA.explained_variance, sklearn.decomposition.PCA.explained_variance_ratio_を用いて取得し、図表等は、numpy.array()を使って、計算した。

ソースコードはプログラムに示す。

プログラム1:主成分分析を行うプログラムのソースコードshc_pca.py

```
'''
主成分分析を行う
'''
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn.decomposition import PCA

def pc_analyse(data: pd.DataFrame, title: str):
    '''
    主成分分析を行う
    Parameters
    -----
    data : pandas.DataFrame
        主成分分析を行うデータ
    name : str
        ... 図表に使うラベル
    '''
    # 標準化
    data_std = data.iloc[:, :].apply(lambda x: (x-x.mean())/x.std(), axis=0)
    # 主成分分析を行う
    pca = PCA()
    pca.fit(data_std)
    feat = pca.transform(data_std)

    # 主成分分析の結果を返す
    score = pd.DataFrame(feat, columns=["PC"+str(i)]
                        for i in range(len(data_std.columns)))
    print("score:\n%score.head()")
    for i in range(len(data_std.columns)):
        score.to_csv("output/"+title+"_score.csv", index=True)

    # 図表、図表、図表をそれぞれ返す
    state = pd.DataFrame([pca.explained_variance_, pca.explained_variance_ratio_, list(np.cumsum(
        pca.explained_variance_ratio_))],
                        columns=["PC"+str(i)] for i in range(len(data_std.columns))],
                        index=["explained_variance", "explained_variance_ratio", "cumulative_sum"])
    print("state:\n%sstate")

    # 図表、図表、図表をそれぞれ返す
    fig = plt.figure(figsize=(5, 5))
    fig = plt.figure(figsize=(5, 5))
    plt.plot(feat[:, 0], feat[:, 1], alpha=0.5, s=list(data_std[:, 0]))
    plt.grid()
    plt.xlabel("PC1")
    plt.ylabel("PC2")
    plt.title("%s(1d, 2d, 3d)" % title)
    plt.show()

    # 図表、図表、図表をそれぞれ返す
    component = pd.DataFrame(pca.components_, index=["PC"+str(i)]
                            for i in range(len(data_std.columns)),
                            columns=data.columns)
    print("component:\n%scomponent.head()")

    component.to_csv("output/"+title+"_component.csv", index=True)
    # 図表、図表、図表をそれぞれ返す
    fig = plt.figure(figsize=(5, 5))
    for x_axis, y_axis, name in zip(pca.components_[0], pca.components_[1], data.columns[1:]):
        plt.plot(x_axis, y_axis, name)
    plt.scatter(pca.components_[0], pca.components_[1], alpha=0.5)

    plt.grid()
    plt.xlabel("PC1")
    plt.ylabel("PC2")
    plt.title("%s(1d, 2d, 3d)" % title)
    plt.show()

    fig.savefig("output/"+title+"_heatmap.png")
```

中学生の成績データの分析

成績のデータはすべて正規分布であるため、すべてのデータを主成分分析に用いた。

実行結果

ソースコード、図表、図表、図表をそれぞれ返すプログラム、図表、図表、図表に示す。

プログラム2:ソースコードseisaku_analyse.py

```
'''
seisaku_1.txtを読み込んで分析する
'''
import pandas as pd

from src.pca import pc_analyse
from src.read_data import read_file

if __name__ == "__main__":
    # データを読み込む
    raw = read_file("output/seisaku_1.txt", "shift-jis")
    data = pc_analyse(raw.iloc[:, :].columns[1:])
    # データを標準化する
    data_std = data.dropna(axis=1, how="any")
    print("data:\n%sdata.head()")
    # 主成分分析を行う
    pc_analyse(data, "score")
```

出力:seisaku_analyse.pyの実行結果

```
5 pipenv run python3 seisaku_analyse.py
data:
  kotoage shakai nogyo rika umakko biyutu taiko gika eigo
0  39.0  43.0  51.0  63.0  68.0  66.0  37.0  44.0  20.0
1  39.0  22.0  48.0  56.0  70.0  72.0  56.0  62.0  16.0
2  39.0  38.0  23.0  57.0  68.0  76.0  33.0  54.0  6.0
3  60.0  47.0  77.0  58.0  77.0  82.0  79.0  66.0  47.0
4  79.0  71.0  76.0  67.0  72.0  82.0  46.0  62.0  44.0
score:
      PC1      PC2      PC3      PC4
0  0.05615  0.36170  0.21702  0.11732
1  0.05615  0.36170  0.21702  0.11732
2  0.05615  0.36170  0.21702  0.11732
3  0.05615  0.36170  0.21702  0.11732
4  0.05615  0.36170  0.21702  0.11732
state:
      PC1      PC2      PC3      PC4
0  0.05615  0.36170  0.21702  0.11732
1  0.05615  0.36170  0.21702  0.11732
2  0.05615  0.36170  0.21702  0.11732
3  0.05615  0.36170  0.21702  0.11732
4  0.05615  0.36170  0.21702  0.11732
component:
  kotoage shakai nogyo ... taiko gika eigo
PC1 -0.36170  0.36170  0.21702  ... 0.11732  0.11732  0.11732
PC2 -0.36170  0.36170  0.21702  ... 0.11732  0.11732  0.11732
PC3 -0.36170  0.36170  0.21702  ... 0.11732  0.11732  0.11732
PC4 -0.36170  0.36170  0.21702  ... 0.11732  0.11732  0.11732
```

表1:主成分分析のデータのデータ

0	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
1	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
2	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
3	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
4	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
5	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
6	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
7	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
8	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
9	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
10	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
11	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
12	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
13	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
14	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
15	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
16	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
17	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
18	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
19	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
20	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
21	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
22	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
23	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
24	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
25	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
26	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
27	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
28	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
29	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
30	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
31	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
32	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
33	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
34	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
35	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
36	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
37	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
38	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
39	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
40	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
41	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
42	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
43	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
44	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
45	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
46	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
47	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
48	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
49	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
50	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
51	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
52	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
53	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
54	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
55	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
56	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
57	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
58	0.147960000000000	0.078800000000000	0.043333333333333	0.050000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000	0.000000000000000
59	0.147960000000000	0.078800000000000	0.0433						

表2: 固有ベクトルのすべてのデータ

また、各生徒と各科目を第一・第二主成分でプロットしたものをそれぞれ図1、図2に示す。

	attack	base_exp_ats	base_happiness	capture_rate	defense	experience_growth	hp	sp_attack	sp_defense	speed
PC1	3.261120864068478	0.235461762051144	0.1152707168317	0.49886040777931937	0.3385415064268286	0.112818134149348	0.3304402482199616	0.3640497993575115	0.384093534487693	0.226594911586248
PC2	5.882440328303775	6.4974747007404965	0.61197496117422	0.897403312947901	0.01381487884328176	0.43433988432131	0.179314748493504	0.1894808080808081	0.24730274181197	0.3303630221105474
PC3	0.3454333333333333	0.02	0.1071707107107107	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.2994784994784995	0.0000000000000000	0.1764860176384444	0.0000000000000000
PC4	0.3105193939393939	0.002295433333333333	0.4048133123472607	0.02670496003333333	0.0714883321551172	0.517649715485405	0.617447462317493	0.1759998938484849	0.2593956074991	0.1988473792846047
PC5	0.6246444444444444	0.5803023333333333	0.01704670810238846	0.01704670810238846	0.01704670810238846	0.01704670810238846	0.27275145435777	0.25938083208083	0.286318847487	0.0000000000000000
PC6	0.2330828282828283	0.618402865476074	0.0011954410515204	0.1864890211430426	0.1868797841926304	0.63481881641325	0.2257641492950676	0.3708546472346244	0.306446766725863	0.69221021737302147
PC7	0.81111838117411	0.2894348059370515	0.4544833194942336	0.10917382983744	0.21382107777668	0.026882821672462	0.273263694378021	0.32981660839768	0.25114444348663	0.40466407584607
PC8	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.2723147710861	0.0000000000000000
PC9	0.34851076401146	0.396467403891285	0.00000000007942	0.87770991401745	0.53063142484974	0.366494955688438	0.16869259770884	0.344870107135653	0.251433664267876	0.654523714773512
PC10	0.438710283644006	0.00027489799964	0.1324362860254038	0.00000000459323	0.000000073363127	0.923411044717533	0.0000000000000000	0.40644883719277	0.298152358988485	0.0000000000000000

Figure 4: A scatter plot titled 'pokemon_distribute' showing the results of a Principal Component Analysis (PCA). The x-axis is labeled 'PC1' and ranges from -4 to 6. The y-axis is labeled 'PC2' and ranges from -6 to 10. The data points are colored by group, with group 1 (purple) and group 2 (teal/green) showing a clear separation along the PC1 axis. There is a small cluster of points at the top right, around PC1=6 and PC2=9.

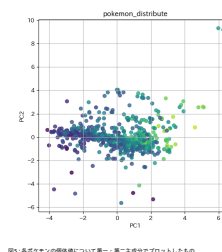


図5:各ボクセンの個体値について第一・第二主成分でプロットしたもの



図6:各個体値について第一・第二主成分でプロットしたもの