|  |  |  |
| --- | --- | --- |
| Лабораторная работа № 4. Метод главных компонент | Студент | Бокова О. Д. |
| Группа | ИВТ-363 |
| Преподаватель | Фокин Р.О. |
| Оценка |  |
| Дата | 03.04.2023 |

1. Выполните анализ главных компонент для выбранного или произвольного набора данных из репозитория UC Irvine Machine Learning

Repository. Выделить главные факторы, приведите интерпретацию результатов (или покажите, что этого сделать невозможно).

2. Выберите два коррелирующих параметра и проведите для них

анализ главных компонент, выделите главные факторы, постарайтесь дать

интепретацию.

3. Сравните результаты по двум пунктам

Код:

# Загрузка необходимых библиотек и датасета

import pandas as pd

import numpy as np

from sklearn.decomposition import PCA

import matplotlib.pyplot as plt

data = pd.read\_csv("biofam.csv", delimiter = ';')

# Проведение предварительного анализа данных и очистки данных

data = data.replace('?', np.nan) # замена пропущенных значений на NaN

data = data.dropna() # удаление строк с пропущенными значениями

# Выбор переменных для PCA

variables = ['a16', 'a20', 'a22', 'a25', 'a27', 'a30']

# Нормализация данных

norm\_data = (data[variables] - data[variables].mean()) / data[variables].std()

# Выполнение PCA

pca = PCA()

pca.fit(norm\_data)

# Выделение главных компонент

components = pca.components\_

scores = pca.transform(norm\_data)

# Производим интерпретацию полученных результатов

print('Доля объясненной дисперсии:', pca.explained\_variance\_ratio\_)

print('Вклад каждой переменной:', components)

# Выбор двух коррелирующих параметров

corr\_variables = ['a16', 'a30']

# Выделение выбранных параметров

corr\_data = norm\_data[corr\_variables]

# Выполнение PCA для выбранных параметров

corr\_pca = PCA()

corr\_pca.fit(corr\_data)

# Выделение главных компонент для выбранных параметров

corr\_components = corr\_pca.components\_

corr\_scores = corr\_pca.transform(corr\_data)

# Производим интерпретацию полученных результатов

print('Доля объясненной дисперсии для выбранных параметров:', corr\_pca.explained\_variance\_ratio\_)

print('Вклад каждой переменной для выбранных параметров:', corr\_components)

# Сравнение результатов

print('Разница в доле объясненной дисперсии:', pca.explained\_variance\_ratio\_.sum() - corr\_pca.explained\_variance\_ratio\_.sum())

# График зависимости доли объясненной дисперсии и количества главных компонент

plt.plot(np.cumsum(pca.explained\_variance\_ratio\_))

plt.xlabel('Number of Components')

plt.ylabel('Variance (%)')

plt.title('Explained Variance by Components')

plt.show()

# График распределения главных компонент

plt.scatter(scores[:, 0], scores[:, 1])

plt.xlabel('1st Principal Component')

plt.ylabel('2nd Principal Component')

plt.title('PCA Biomedical Family Data')

plt.show()

Результат:

