Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Кафедра комп'ютерного моделювання процесів і систем

ЗВІТ

з лабораторної роботи №11

"Зниження розмірності за допомогою PCA та SVD"

з курсу

«Алгоритми та моделі збору, аналізу та візуалізації даних»

Виконав: студент групи ІКМ-М222к Черкас Ю.В.

Перевірила: аспірантка Рикова В.О.

Харків 2023р

Варіант №15

В роботі для зниження розмірності використовується бібліотека scikit-learn https://scikit-learn.org/stable/modules/manifold.html

Виконання

Для відповідного датасету згідно з варіантом виконати пониження розмірності даних за допомогою PCA та SVD. Датасети розміщенні в теці datasets (https://github.com/a-vodka/dv/tree/master/lab/dataset).

1. Використовуючи PCA візуалізувати данні у просторах з розмірностями два та три (2D та 3D).

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.decomposition import PCA

data = pd.read_csv('Wholesale customers data.csv')
X = data.iloc[:,:-1].values
y = data.iloc[:,-1].values

pca_2d = PCA(n_components=2)
pca_3d = PCA(n_components=3)
X_2d = pca_2d.fit_transform(X)
X_3d = pca_3d.fit_transform(X)

x_min, x_max = np.min(X_2d), np.max(X_2d)
X_2d = (X_2d - x_min) / (x_max - x_min)
x_min, x_max = np.min(X_3d), np.max(X_3d)
X_3d = (X_3d - x_min) / (x_max - x_min)

plt.figure()
plt.scatter(X_2d[:,0], X_2d[:,1], c = y)
plt.title('PCA 2D')
plt.figure()
plt.subplot(111, projection = '3d')
plt.scatter(X_3d[:,0], X_3d[:,1], X_3d[:,2], c = y)
plt.title('PCA 3D')
plt.show()
```

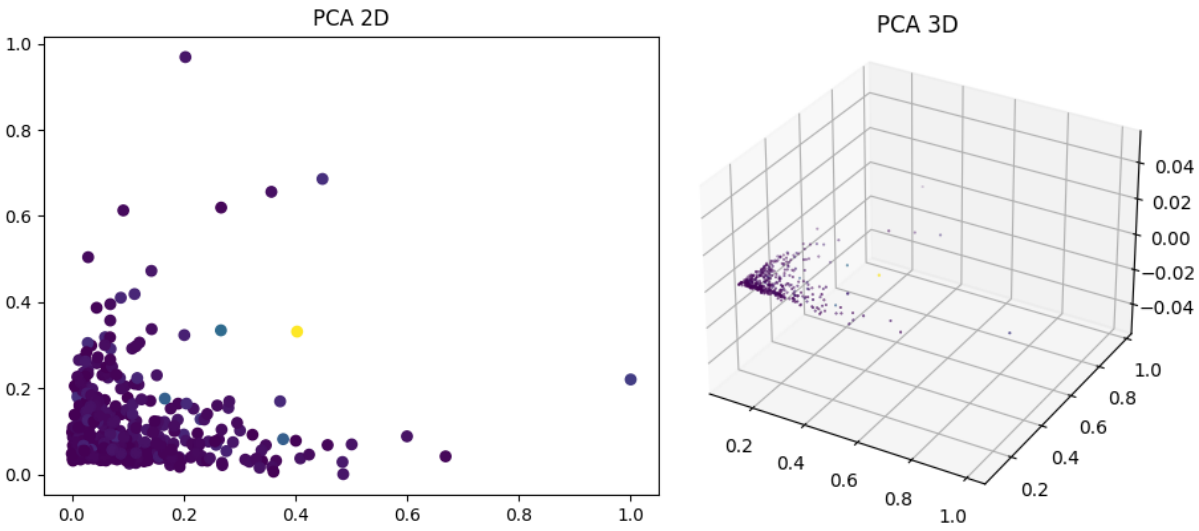

Рисунок 1 – 2D та 3D візуалізація даних за допомогою PCA

2. Використовуючи SVD, побудувати графік залежності власних значень матриці від їх номеру. Перед побудовою графіку впорядкувати власні значення у спадаючому порядку.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import TruncatedSVD

data = pd.read_csv('Wholesale customers data.csv')

svd = TruncatedSVD(n_components=data.shape[1])
svd.fit(data)

own_values = svd.singular_values_
idx_sorted = np.argsort(own_values)[::-1]
own_values_sorted = own_values[idx_sorted]

plt.plot(own_values_sorted, marker="*")
plt.xlabel('Index')
plt.ylabel('Own Value')
plt.grid(True)
plt.show()
```

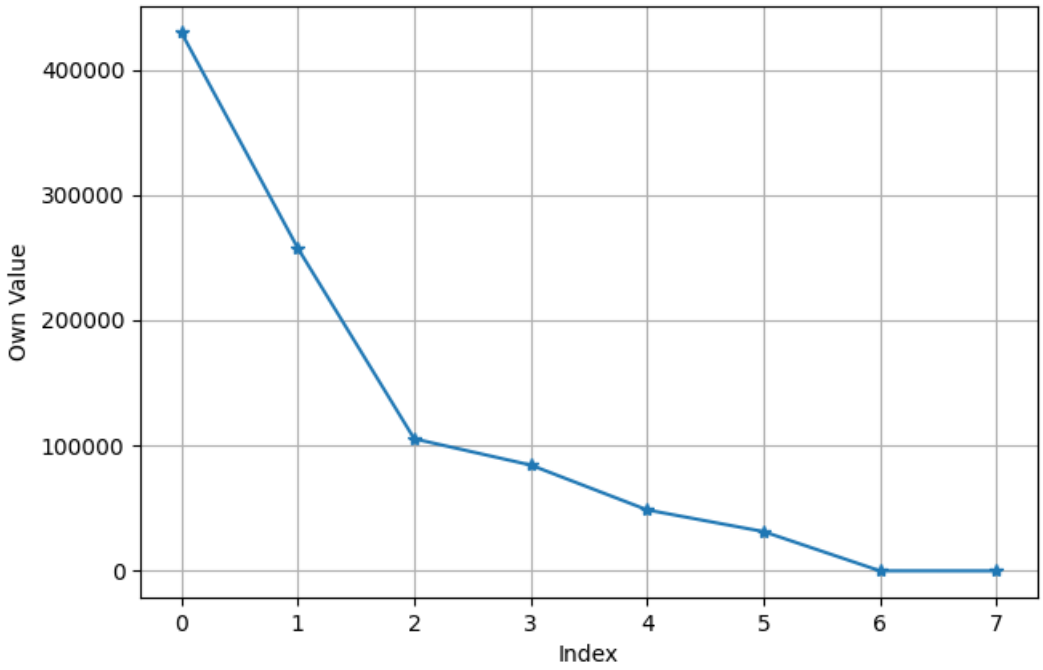

Рисунок 2 – Графік залежності власних значень матриці від їх номеру

3. Визначити таке найменше значення розміру простору d, для якого виконується співвідношення (1). Де λ_i – власні значення матриці, n – загальна кількість власних значень.

$$\frac{\sum_{i=0}^{d} \lambda_i}{\sum_{i=0}^{n} \lambda_i} \leq 0.8 \qquad (1)$$

4. Занулити λ_i, для яких $d \leq i \leq n$. Виконати зворотне перетворення та порівняти отримані данні з вихідними. За можливості побудувати візуалізацію отриманих даних після зворотного перетворення.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import TruncatedSVD
```

```
data = pd.read_csv('Wholesale customers data.csv').iloc[:,2:].values

svd = TruncatedSVD(n_components=data.shape[1]-1)
X_svd = svd.fit_transform(data)
own_values = svd.singular_values_

total = np.sum(own_values)
target_dimenstion_size = 0

for i in range(len(own_values)):
    current_sum = np.sum(own_values[:i+1])
    ratio = current_sum / total
    if ratio >= 0.8:
        break
    target_dimenstion_size = i

print('Target dimenstion size: ', i)

own_values[target_dimenstion_size-1:] = 0
svd.singular_values_ = own_values

data_inverse = svd.inverse_transform(X_svd)

np.set_printoptions(precision=1)
print('Original dataset:')
print(data[:3, :])
print('Restored dataset:')
print(data_inverse[:3, :])

plt.figure()
plt.plot(data)
plt.figure()
plt.plot(data_inverse)
plt.show()
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL

Target dimenstion size:  2
Original dataset:
[[12669  9656  7561   214  2674  1338]
 [ 7057  9810  9568  1762  3293  1776]
 [ 6353  8808  7684  2405  3516  7844]]
Restored dataset:
[[12677.5  9643.8  7413.4   222.4  2994.9  1436.4]
 [ 7071.3  9789.5  9320.4  1776.1  3831.2  1941. ]
 [ 6306.7  8874.6  8486.4  2359.4  1771.6  7309.1]]
```

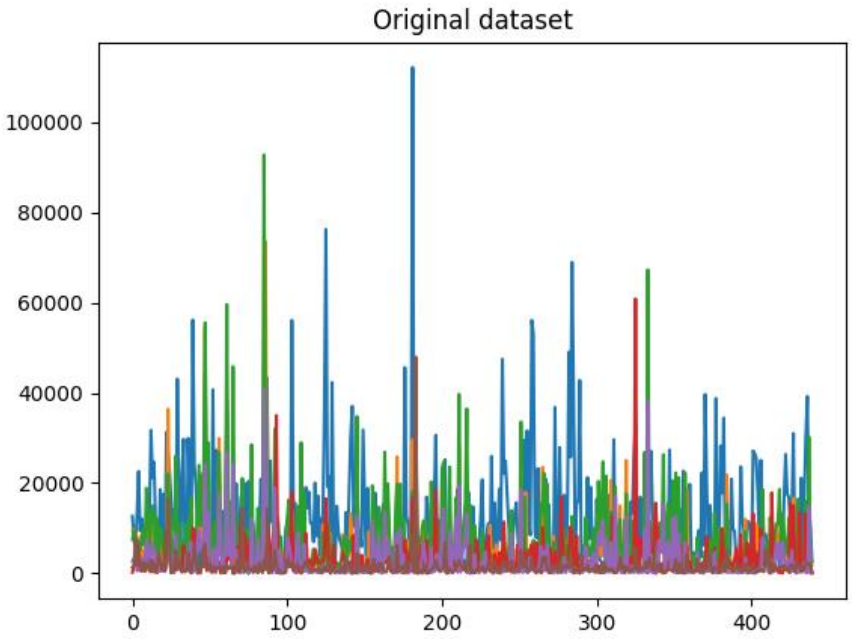

Рисунок 3 – Графік рядів значень початкового набору даних

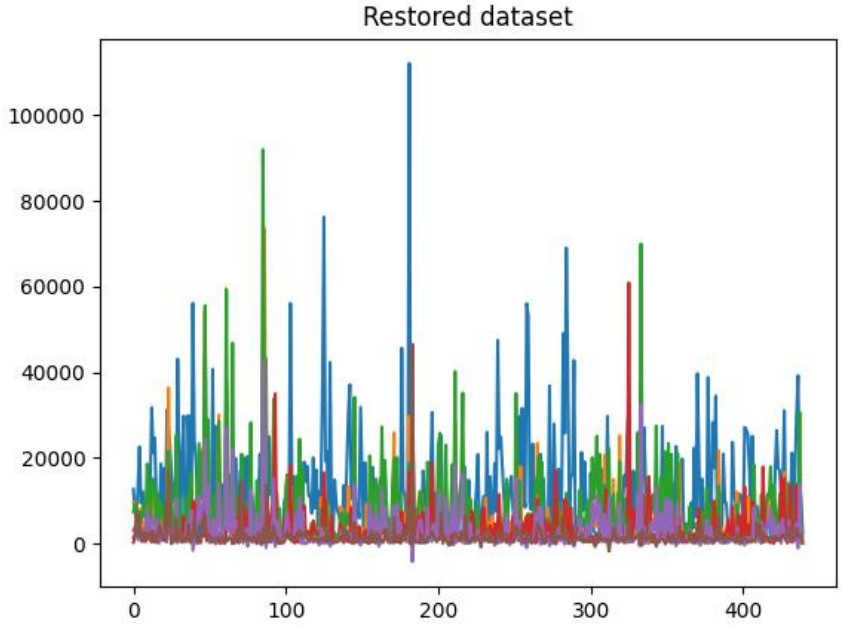

Рисунок 4 – Графік рядів значень відновленого набору даних

Висновок: на даній лабораторній роботі ми провели аналіз даних високої розмірності за допомогою PCA та SVD. Навчилися будувати графік залежності власних значень матриці від їх номеру. Дослідили вплив власних значень матриці при зворотному перетворені даних, а саме занулення тих власних значень, котрі мають мінімальний вплив.