Липецкий государственный технический университет

Факультет автоматизации и информатики

Кафедра автоматизированных систем управления

ЛАБОРАТОРНАЯ РАБОТА №4

по дисциплине «Прикладные интеллектуальные системы и экспертные системы»

Кластеризация данных

Студент Коровайцева А.В.

Группа М-ИАП-23-1

Руководитель Кургасов В.В.

Доцент

Липецк 2023 г.

Цель работы

Получить практические навыки решения задачи кластеризации фактографических данных в среде Jupiter Notebook. Научиться настраивать параметры методов и оценивать точность полученного разбиения.

Задание кафедры

1) Загрузить выборки согласно варианту задания.

2) Отобразить данные на графике в пространстве признаков. Поскольку решается задача кластеризации, то подразумевается, что априорная информация о принадлежности каждого объекта истинному классу неизвестна, соответственно, на данном этапе все объекты на графике должны отображаться одним цветом, без привязки к классу.

3) Провести иерархическую кластеризацию выборки, используя разные способы вычисления расстояния между кластерами: расстояние ближайшего соседа (single), дальнего соседа (complete), Уорда (Ward). Построить дендрограммы для каждого способа. Размер графика должен быть подобран таким образом, чтобы дендрограмма хорошо читалась.

4) Исходя из дендрограмм выбрать лучший способ вычисления расстояния между кластерами.

5) Для выбранного способа, исходя из дендрограммы, определить количество кластеров в имеющейся выборке. Отобразить разбиение на кластеры и центроиды на графике в пространстве признаков (объекты одного кластера должны отображаться одним и тем же цветом, центроиды всех кластеров – также одним цветом, отличным от цвета кластеров).

6) Рассчитать среднюю сумму квадратов расстояний до центроида, среднюю сумму межкластерных расстояний для данного разбиения. Сделать вывод о качестве разбиения.

7) Провести кластеризацию выборки методом k-средних. для k [1, 10].

8) Сформировать три графика: зависимость средней суммы квадратов расстояний до центроида, средней суммы средних внутрикластерных расстояний и средней суммы межкластерных расстояний от количества кластеров. Исходя из результатов, выбрать оптимальное количество кластеров.

9) Составить сравнительную таблицу результатов разбиения иерархическим методом и методом k-средних.

Вариант №8

n_samples = 100

Вид классов: classification

random_state = 36

class_sep = 1.2

Для всех вариантов:

n_features = 2

n_redundant = 0

n_informative = 2

n_cluster_per_class = 1

n_classes = 4

Ход работы

Загрузим выборки по варианту. Код для генерации данных представлен на рисунке 1.

```
X, y = make_classification(n_samples=100,
                           n_features=2,
                           n_redundant=0,
                           n_informative=2,
                           n_clusters_per_class=1,
                           n_classes=4,
                           random_state=36,
                           class_sep=1.2)
```

Рисунок 1 – Код для генерации данных

Отобразим на графике сгенерированные данные, для этого воспользуемся библиотекой matplotlib.pyplot. Полученный график представлен на рисунке 2.

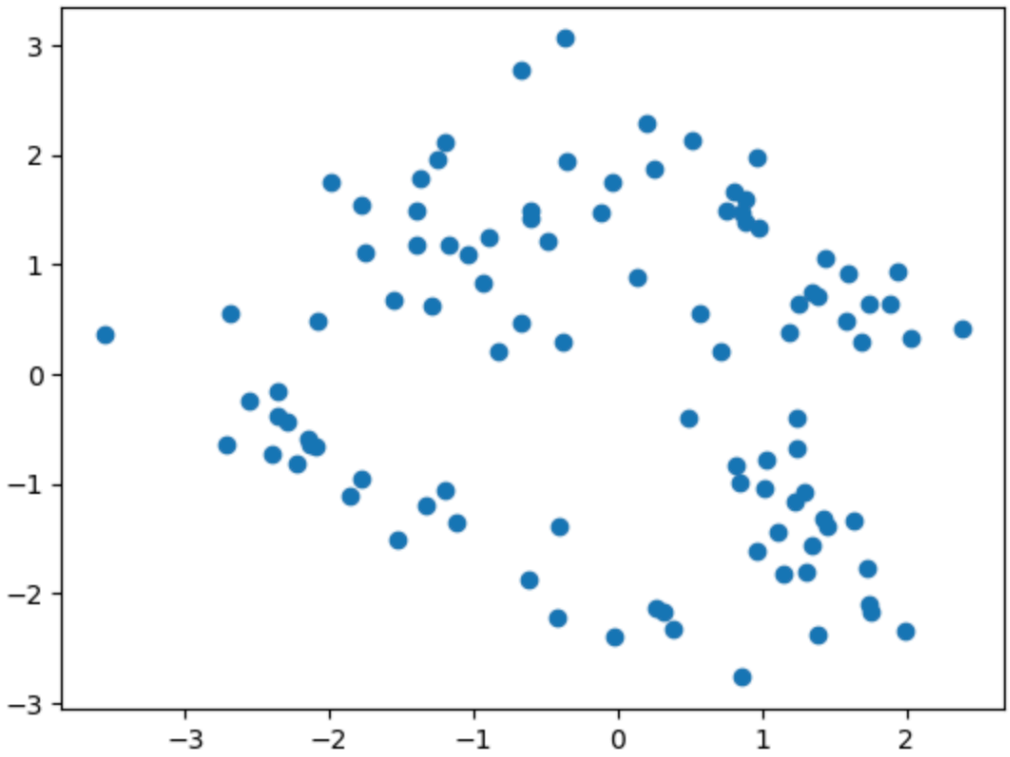

Рисунок 2 – Визуализация сгенерированных данных

Воспользуемся иерархической кластеризацией выборки с использованием различных методов вычисления расстояния.

Метод вычисления расстояния ближайшего соседа (single) и код для этого представлен на рисунке 3. Полученная дендограмма представлена на рисунке 4.

```
mergings_single = linkage(X, method='single')
mergings_single
```

```
ay([[1.60000000e+01, 7.90000000e+01, 4.61700718e-02, 2.00000000e+00],
    [5.90000000e+01, 1.00000000e+02, 4.82382520e-02, 3.00000000e+00],
    [4.00000000e+01, 7.00000000e+01, 5.11405618e-02, 2.00000000e+00],
    [1.20000000e+01, 2.60000000e+01, 5.78546254e-02, 2.00000000e+00],
    [1.30000000e+01, 5.60000000e+01, 6.88632557e-02, 2.00000000e+00],
    [3.30000000e+01, 6.20000000e+01, 7.01583969e-02, 2.00000000e+00],
    [1.70000000e+01, 3.00000000e+01, 7.86585329e-02, 2.00000000e+00],
    [5.00000000e+01, 5.50000000e+01, 8.59038033e-02, 2.00000000e+00],
    [8.00000000e+00, 3.60000000e+01, 8.74654507e-02, 2.00000000e+00],
    [8.00000000e+01, 1.07000000e+02, 9.82539902e-02, 3.00000000e+00],
    [3.80000000e+01, 1.09000000e+02, 1.04612512e-01, 4.00000000e+00],
    [4.90000000e+01, 8.90000000e+01, 1.07241587e-01, 2.00000000e+00],
```

Рисунок 3 – Вычисленные расстояния ближайшего соседа (single)

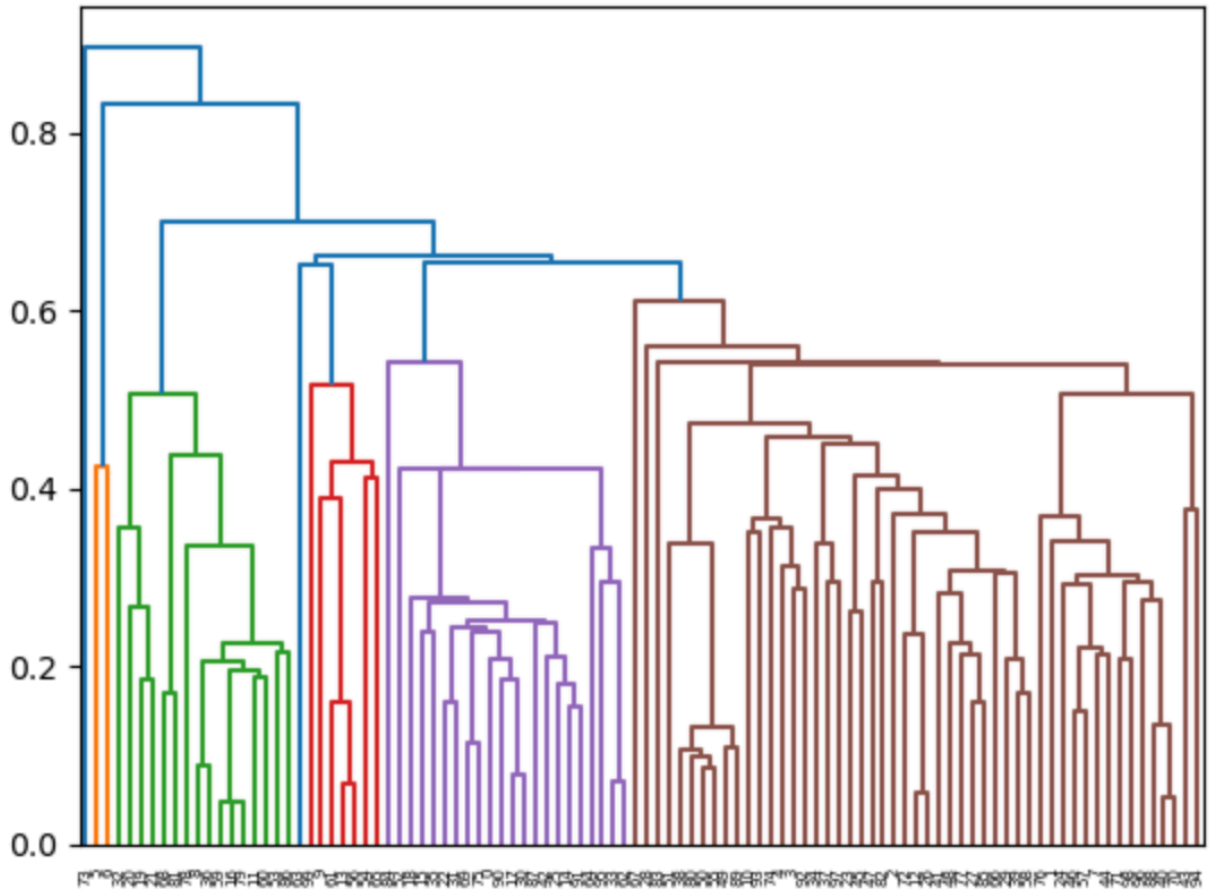

Рисунок 4 – Дендограмма для расстояния ближайшего соседа (single)

Метод вычисления расстояния дальнего соседа (complete) и код для этого представлен на рисунке 5. Полученная дендограмма представлена на рисунке 6.

```
mergings_complete = linkage(X, method='complete')
mergings_complete
```

```
ay([[1.60000000e+01, 7.90000000e+01, 4.61700718e-02, 2.00000000e+00],
    [4.00000000e+01, 7.00000000e+01, 5.11405618e-02, 2.00000000e+00],
    [1.20000000e+01, 2.60000000e+01, 5.78546254e-02, 2.00000000e+00],
    [1.30000000e+01, 5.60000000e+01, 6.88632557e-02, 2.00000000e+00],
    [3.30000000e+01, 6.20000000e+01, 7.01583969e-02, 2.00000000e+00],
    [1.70000000e+01, 3.00000000e+01, 7.86585329e-02, 2.00000000e+00],
    [5.00000000e+01, 5.50000000e+01, 8.59038033e-02, 2.00000000e+00],
    [8.00000000e+00, 3.60000000e+01, 8.74654507e-02, 2.00000000e+00],
    [5.90000000e+01, 1.00000000e+02, 9.07472797e-02, 3.00000000e+00],
    [4.90000000e+01, 8.90000000e+01, 1.07241587e-01, 2.00000000e+00],
    [6.90000000e+01, 7.50000000e+01, 1.13102713e-01, 2.00000000e+00],
    [8.80000000e+01, 1.01000000e+02, 1.34292668e-01, 3.00000000e+00],
```

Рисунок 5 – Вычисленные расстояния дальнего соседа (complete)

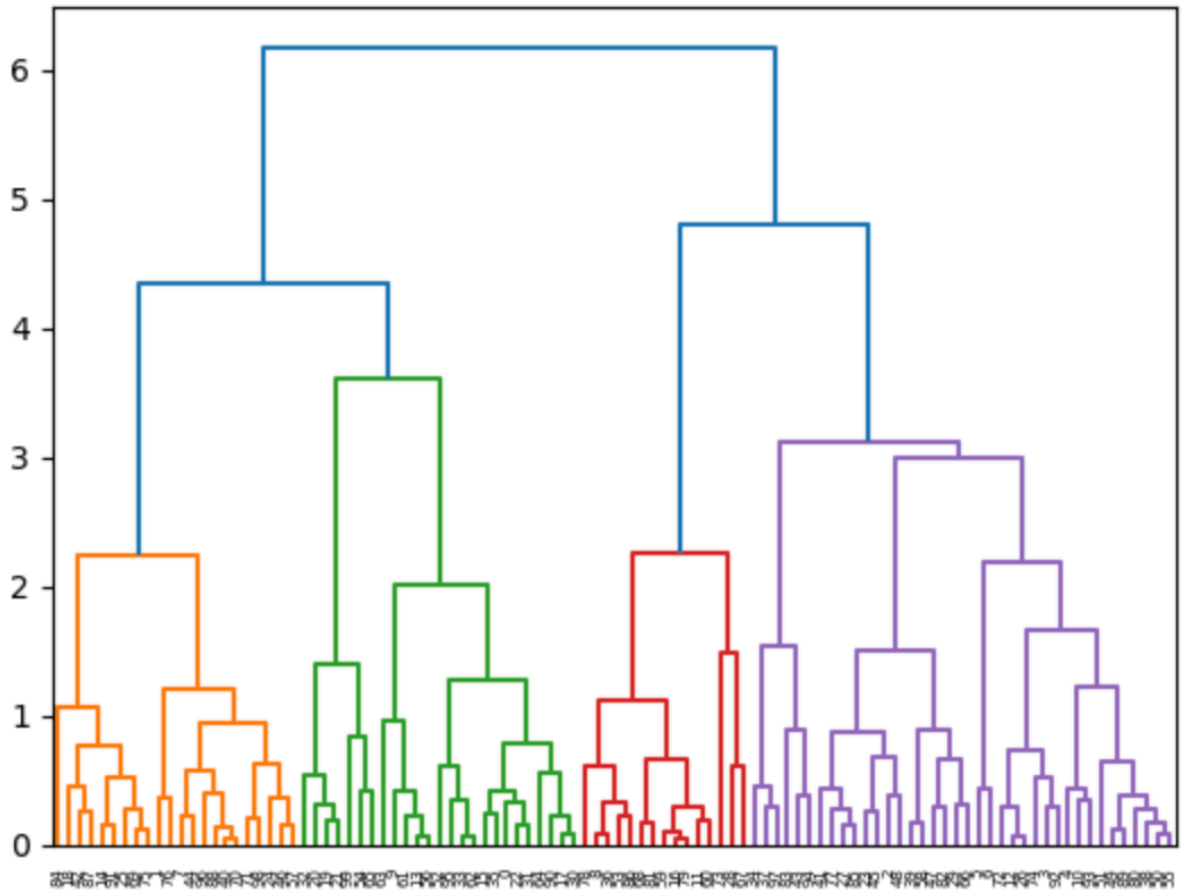

Рисунок 6 – Дендограмма для расстояния дальнего соседа (complete)

Метод вычисления расстояния Уорда (ward) и код для этого представлен на рисунке 7. Полученная дендограмма представлена на рисунке 8.

```
mergings_ward = linkage(X, method='ward')
mergings_ward
```

```
ay([[1.60000000e+01, 7.90000000e+01, 4.61700718e-02, 2.00000000e+00],
    [4.00000000e+01, 7.00000000e+01, 5.11405618e-02, 2.00000000e+00],
    [1.20000000e+01, 2.60000000e+01, 5.78546254e-02, 2.00000000e+00],
    [1.30000000e+01, 5.60000000e+01, 6.88632557e-02, 2.00000000e+00],
    [3.30000000e+01, 6.20000000e+01, 7.01583969e-02, 2.00000000e+00],
    [1.70000000e+01, 3.00000000e+01, 7.86585329e-02, 2.00000000e+00],
    [5.90000000e+01, 1.00000000e+02, 7.95661568e-02, 3.00000000e+00],
    [5.00000000e+01, 5.50000000e+01, 8.59038033e-02, 2.00000000e+00],
    [8.00000000e+00, 3.60000000e+01, 8.74654507e-02, 2.00000000e+00],
    [4.90000000e+01, 8.90000000e+01, 1.07241587e-01, 2.00000000e+00],
    [6.90000000e+01, 7.50000000e+01, 1.13102713e-01, 2.00000000e+00],
    [4.60000000e+01, 5.70000000e+01, 1.48459346e-01, 2.00000000e+00],
```

Рисунок 7 – Вычисленные расстояния с помощью метода Уорда (ward)

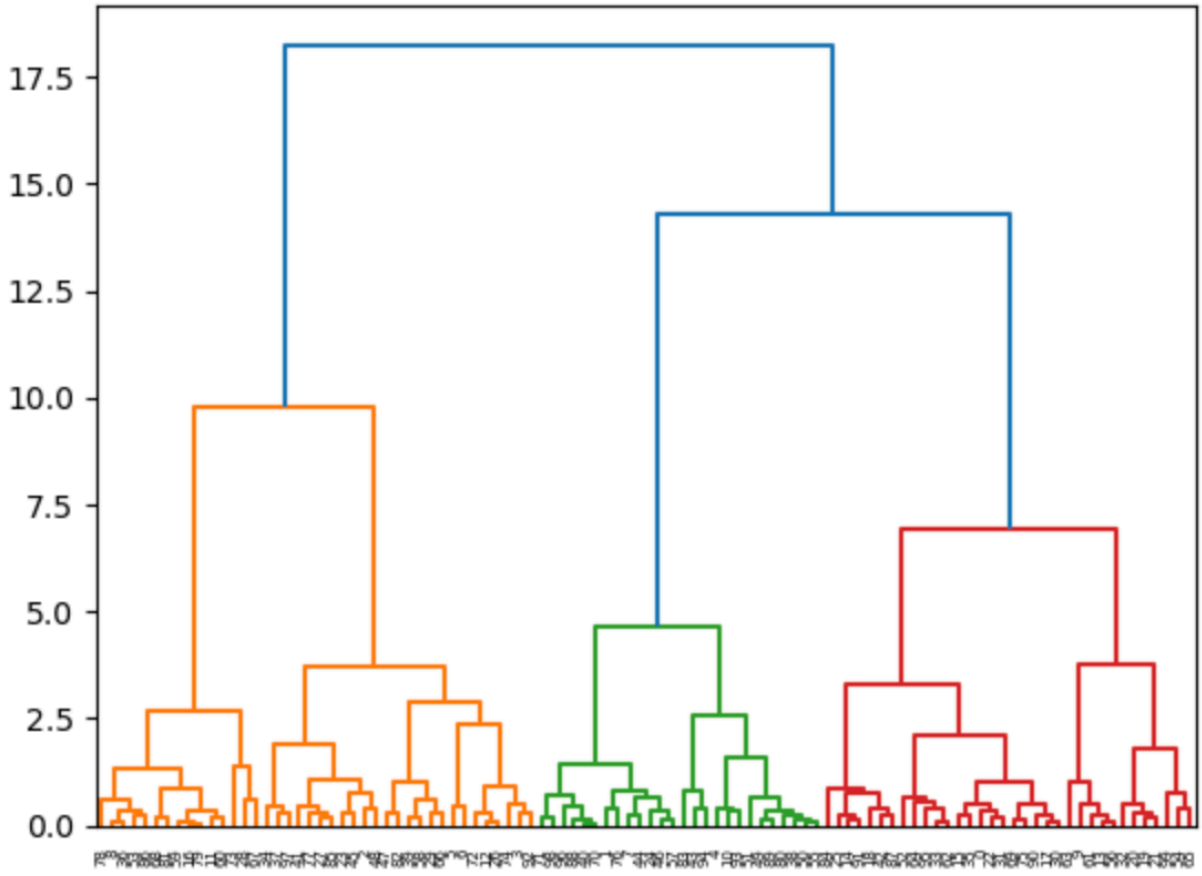

Рисунок 8 – Дендограмма для расстояния Уорда (ward)

Лучшим способом вычисления расстояния между кластерами является дальнего соседа (complete). Определим количество кластеров в имеющейся выборке с использованием данного способа и отобразим разбиение на

кластеры и центроиды на графике в пространстве признаков. Полученное разбиение представлено на рисунке 9.

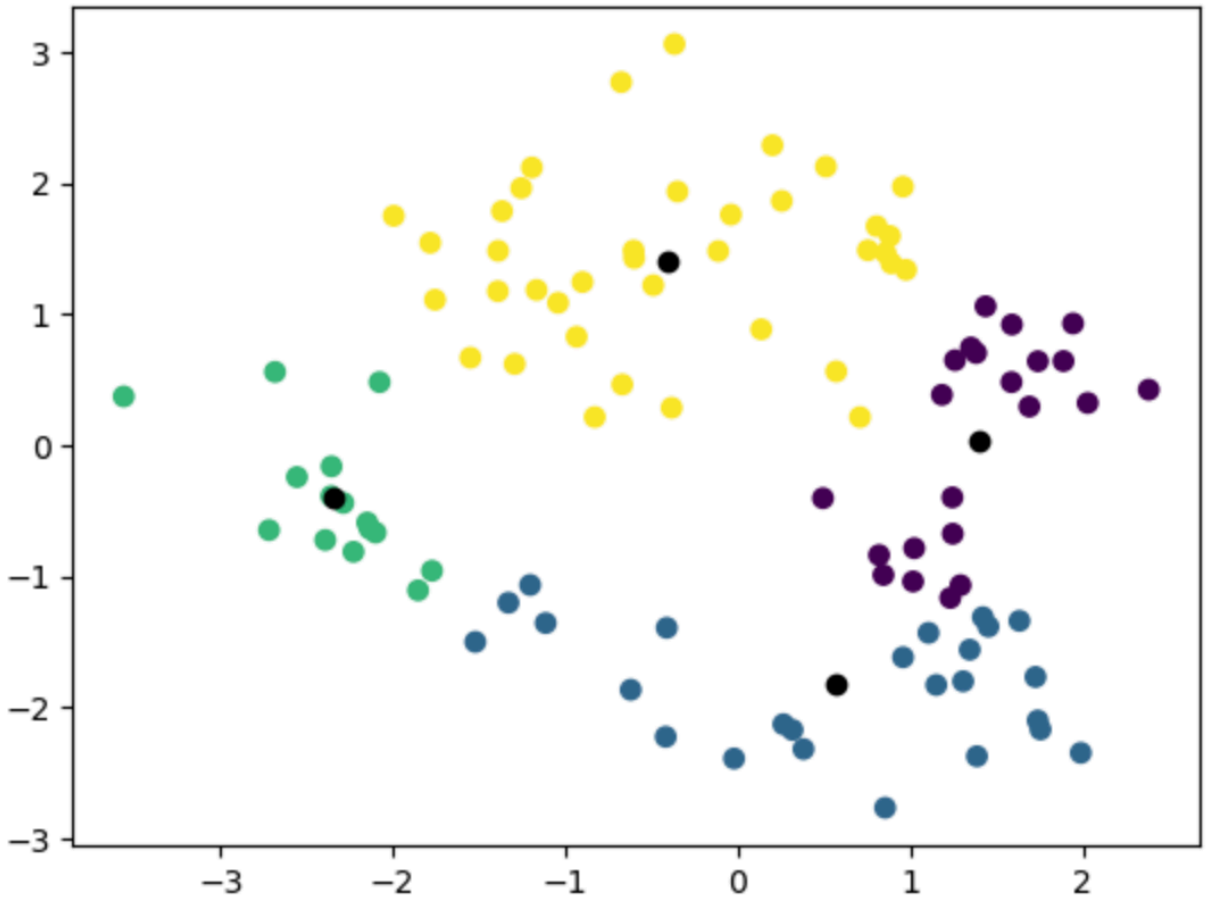

Рисунок 9 – График разбиения данных на кластеры

Рассчитаем среднюю сумму квадратов расстояний до центроида, среднюю сумму средних внутрикластерных расстояний и среднюю сумму межкластерных расстояний для данного разбиения.

Рассчитанное значение суммы квадратов расстояний до центроида и код для этого представлен на рисунке 10. Рассчитанное значение средних внутрикластерных расстояний и код для этого представлен на рисунке 11. Рассчитанное значение суммы межкластерных расстояний и код для этого представлен на рисунке 12.

```
sum_sq_dist = np.zeros(4)
for i in range(1, 5):
    ix = np.where(T == i)
    sum_sq_dist[i - 1] = np.sum(euclidean_distances(*X[ix, :], [clusters[i - 1]]) ** 2)
sum_sq_dist = np.sum(sum_sq_dist) / 4
sum_sq_dist
```

```
25.843203994463714
```

Рисунок 10 – Сумма квадратов расстояний до центроида

```
sum_avg_intercluster_dist = np.zeros(4)
for i in range(1, 5):
    ix = np.where(T == i)
    sum_avg_intercluster_dist[i - 1] = np.sum(euclidean_distances(*X[ix, :], [clusters[i - 1]]) ** 2) / len(*X[i
sum_avg_intercluster_dist = np.sum(sum_avg_intercluster_dist) / 4
sum_avg_intercluster_dist
```

```
0.9399837208674937
```

Рисунок 11 – Сумма средних внутрикластерных расстояний

```
sum_intercluster_dist = np.sum(euclidean_distances(clusters, clusters))
sum_intercluster_dist
```

```
34.64882792358676
```

Рисунок 12 – Сумма межкластерных расстояний

Проведем кластеризацию выборки методом k-средних для k = [1, 10], а после построим три графика: зависимость средней суммы квадратов расстояний до центроида, средней суммы средних внутрикластерных расстояний и средней суммы межкластерных расстояний от количества кластеров.

Код для расчета средней суммы квадратов расстояний до центроида представлен на рисунке 13, а построенный график на рисунке 14.

```
sum_sq_dist_avg = []
for it, kmean in enumerate(models):
    sum_sq_dist_avg.append(kmean.inertia_ / (it + 1))
sum_sq_dist_avg
```

```
[403.07111622330194,
 110.30529486555724,
 40.34514738171166,
 18.42231088866526,
 11.920247809909446,
 7.940148267162104,
 5.895817325260838,
 4.51018103443448,
 3.291952979883026,
 2.5322764756413925]
```

Рисунок 13 – Код для вычисления средней суммы квадратов расстояний до центроида

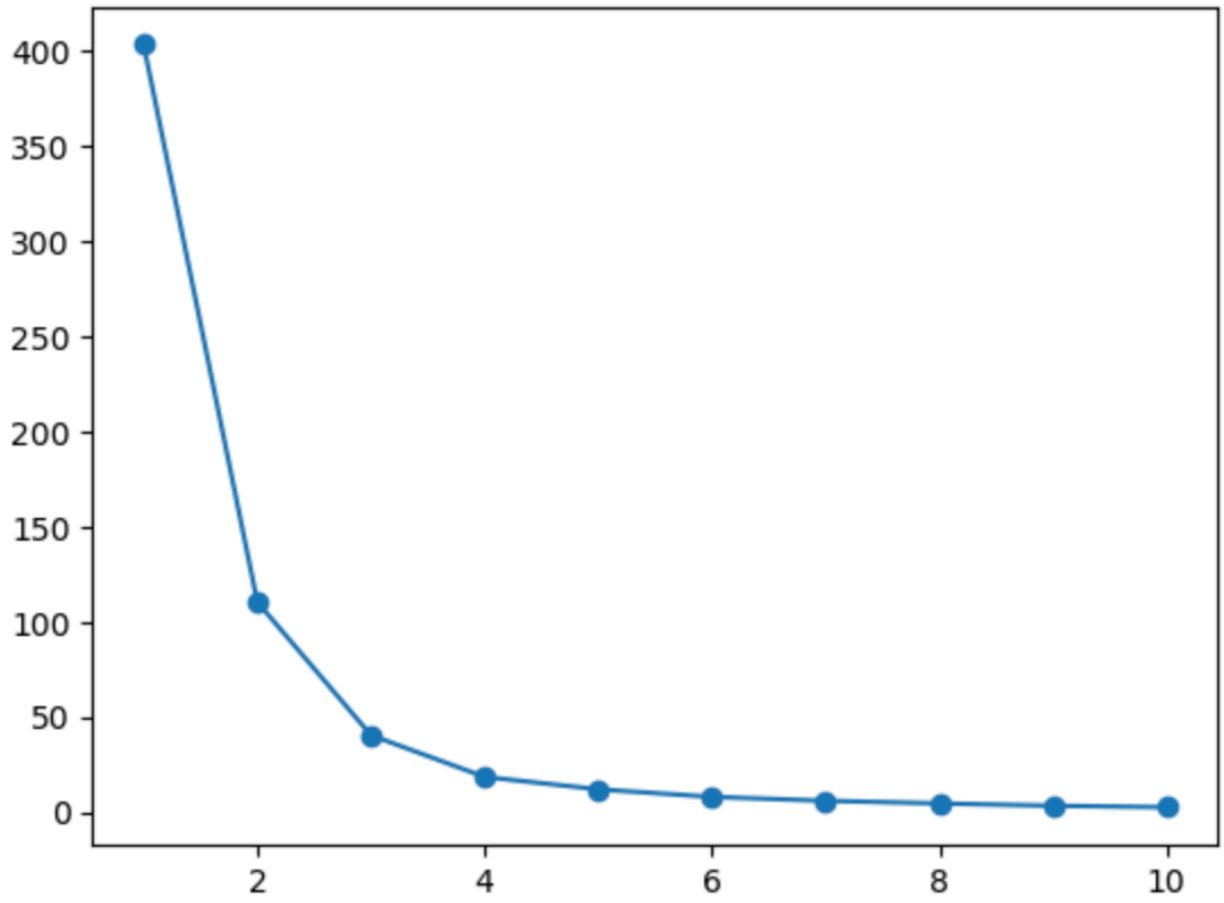

Рисунок 14 – График средней суммы квадратов расстояний до центроида

Код для расчета средней суммы средних внутрикластерных расстояний представлен на рисунке 15, а построенный график на рисунке 16.

```
new_centers = [kmean.cluster_centers_ for kmean in models]

sum_avg_intercluster_dist_avg = []
for k, kmean in enumerate(models):
    intercluster_sum = np.zeros(4)
    for i in range(4):
        ix = np.where(predicted_values[k] == i)
        if len(ix[0]) == 0:
            intercluster_sum[i - 1] = 0
        else:
            intercluster_sum[i - 1] = np.sum(euclidean_distances(*X[ix, :], [kmean.cluster_centers_[i - 1]]) **
    sum_avg_intercluster_dist_avg.append(np.sum(intercluster_sum) / (k + 1))
sum_avg_intercluster_dist_avg
```

```
[4.030711162233019,
 9.539818208956952,
 9.722873156214417,
 10.299566956099865,
 8.391771367689321,
 6.249375960463264,
 5.445107229284192,
 4.677511901785664,
 7.05546172286521,
 2.649105068258345]
```

Рисунок 15 – Код для вычисления средней суммы средних внутрикластерных расстояний

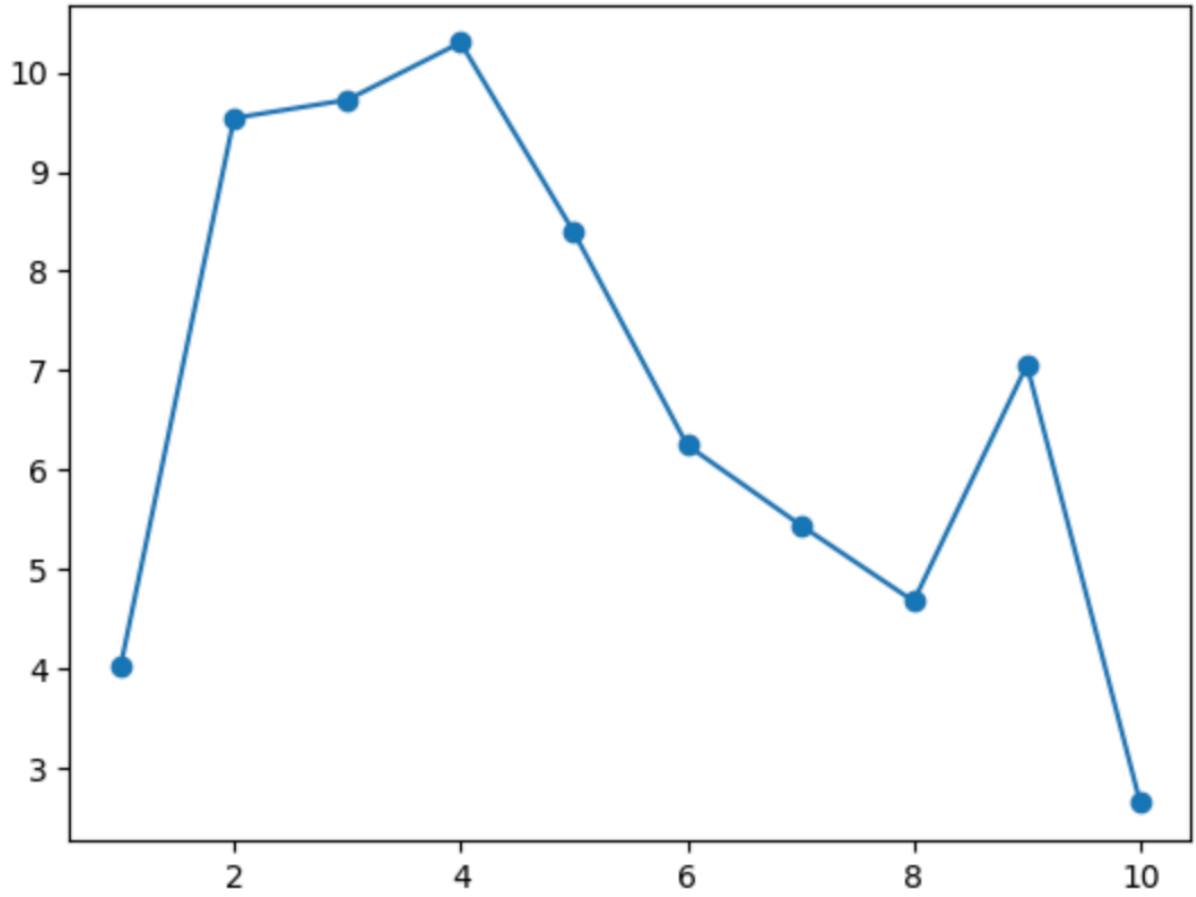

Рисунок 16 – График средней суммы средних внутрикластерных расстояний

Код для расчета средней суммы межкластерных расстояний от количества кластеров представлен на рисунке 17, а построенный график на рисунке 18.

```
sum_intercluster_dist_avg = []

for k, kmean in enumerate(models):
    value = np.sum(euclidean_distances(kmean.cluster_centers_, kmean.cluster_centers_))
    sum_intercluster_dist_avg.append(value / (k + 1))
sum_intercluster_dist_avg
```

```
[0.0,
 2.7151689605865834,
 5.824907427629434,
 8.778857420489008,
 11.111796453200029,
 14.073958288905624,
 16.333348130911535,
 19.206850733209073,
 22.532207165164277,
 25.448612512468152]
```

Рисунок 17 – Код для вычисления средней суммы межкластерных расстояний от количества кластеров

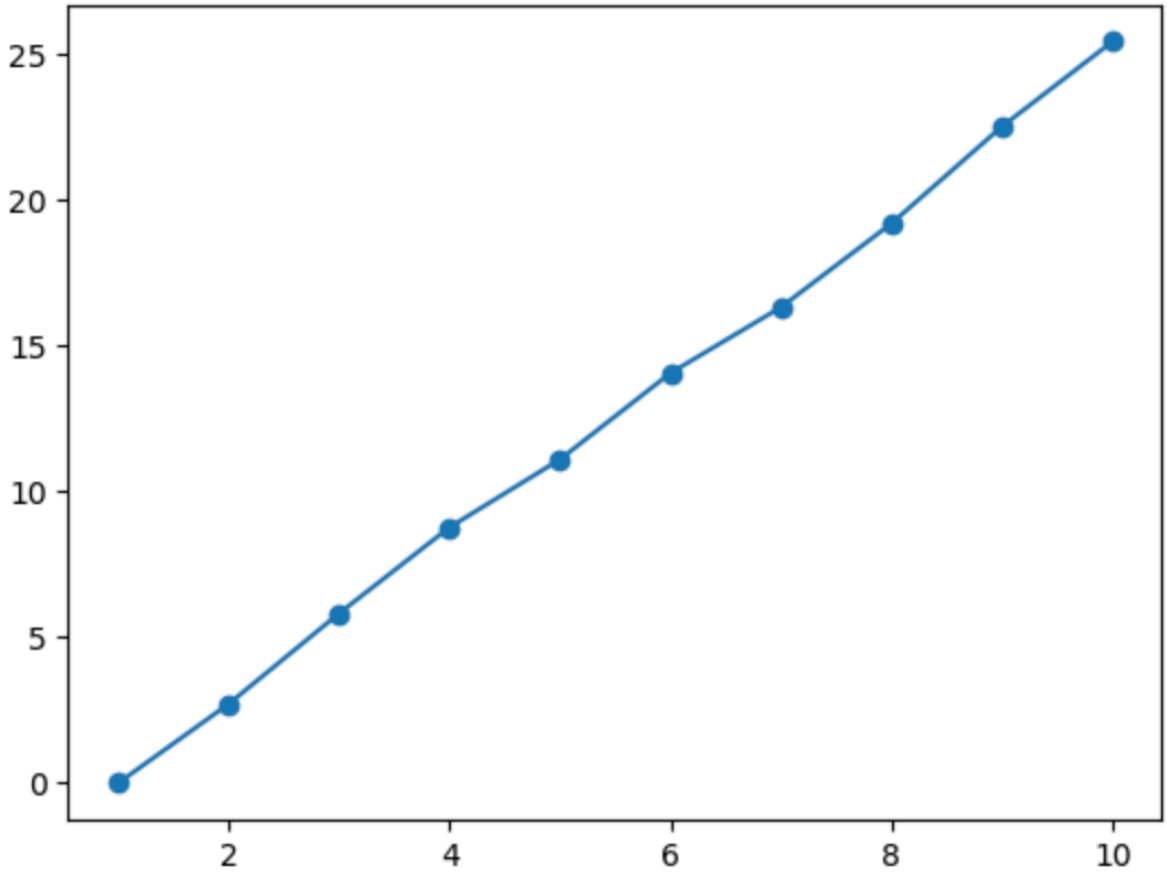

Рисунок 18 – График средней суммы межкластерных расстояний от количества кластеров

Как видно из рисунка 14 оптимальное количество кластеров составляет от двух до четырех.

Составим сравнительную таблицу для ранее описанных метрик качества моделей для иерархического метода и метода k-средних. Составленная таблица представлена на рисунках 19 – 21.

	Иерархический метод			Метод k-средних		
	Сумма квадратов расстояний до центроида	Сумма средних внутрикластерных расстояний	Сумма межкластерных расстояний	Сумма квадратов расстояний до центроида	Сумма средних внутрикластерных расстояний	Сумма межкластерных расстояний
0	25,84320399	0,939983721	34,64882792	403,0711162	4,030711162	0
1	25,84320399	0,939983721	34,64882792	110,3052949	9,539818209	2,715168961
2	25,84320399	0,939983721	34,64882792	40,34514738	9,722873156	5,824907428
3	25,84320399	0,939983721	34,64882792	18,42231089	10,29956696	8,77885742
4	25,84320399	0,939983721	34,64882792	11,92024781	8,391771368	11,11179645
5	25,84320399	0,939983721	34,64882792	7,940148267	6,24937596	14,07395829
6	25,84320399	0,939983721	34,64882792	5,895817325	5,445107229	16,33334813
7	25,84320399	0,939983721	34,64882792	4,510181034	4,677511902	19,20685073
8	25,84320399	0,939983721	34,64882792	3,29195298	7,055461723	22,53220717
9	25,84320399	0,939983721	34,64882792	2,532276476	2,649105068	25,44861251

Рисунок 19 – Сводная таблица сравнения

Иерархический метод		
Сумма квадратов расстояний до центроида	Сумма средних внутрикластерных расстояний	Сумма межкластерных расстояний
25,84320399	0,939983721	34,64882792
25,84320399	0,939983721	34,64882792
25,84320399	0,939983721	34,64882792
25,84320399	0,939983721	34,64882792
25,84320399	0,939983721	34,64882792
25,84320399	0,939983721	34,64882792
25,84320399	0,939983721	34,64882792
25,84320399	0,939983721	34,64882792
25,84320399	0,939983721	34,64882792
25,84320399	0,939983721	34,64882792

Рисунок 20 – Таблица данных для иерархического метода

Метод k-средних		
Сумма квадратов расстояний до центроида	Сумма средних внутрикластерных расстояний	Сумма межкластерных расстояний
403,0711162	4,030711162	0
110,3052949	9,539818209	2,715168961
40,34514738	9,722873156	5,824907428
18,42231089	10,29956696	8,77885742
11,92024781	8,391771368	11,11179645
7,940148267	6,24937596	14,07395829
5,895817325	5,445107229	16,33334813
4,510181034	4,677511902	19,20685073
3,29195298	7,055461723	22,53220717
2,532276476	2,649105068	25,44861251

Рисунок 21 – Таблица данных для метода k-средних

Вывод

В ходе выполнения данной лабораторной работы мною были получены навыки кластеризации данных. В рамках данной работы были применены различные методы кластеризации: иерархический метод и метод k-средних.

В ходе анализа метрик было определено, что оптимальное значение кластеров от 2 до 4. Также была составлена таблица сравнения метрик иерархическим методом кластеризации и методом k-средних.

Приложение А

Исходный код

```
#!/usr/bin/env python
# coding: utf-8

# # Лабораторная работа №4
# ### Вариант №8
#
# ### Вид классов: `classification`
# ### Random state: `36`
# ### Class sep: `1.2`
#
# ### Для всех:
# ### `n_features = 2`
# ### `n_redundant = 0`
# ### `n_informative = 2`
# ### `n_clusters_per_class = 1`
# ### `n_classes = 4`
# ### `n_samples = 100`

# In[1]:


from sklearn.datasets import make_classification


# ### Загрузка выборки согласно варианту №7

# In[2]:


X, y = make_classification(n_samples=100,
                           n_features=2,
                           n_redundant=0,
                           n_informative=2,
                           n_clusters_per_class=1,
                           n_classes=4,
                           random_state=36,
                           class_sep=1.2)


# ### Отображение выборки на графике

# In[3]:


import matplotlib.pyplot as plt


# In[4]:


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


# ### Иерархическая кластеризация выборки

# In[5]:


from scipy.cluster.hierarchy import linkage, dendrogram
```

```
# ### Расстояние ближайшего соседа (single)

# In[6]:


mergings_single = linkage(X, method='single')
mergings_single


# In[7]:


dendrogram(mergings_single)
plt.show()


# ### Расстояние дальнего соседа (complete)

# In[8]:


mergings_complete = linkage(X, method='complete')
mergings_complete


# In[9]:


dendrogram(mergings_complete)
plt.show()


# ### Расстояние Уорда (Ward)

# In[10]:


mergings_ward = linkage(X, method='ward')
mergings_ward


# In[11]:


dendrogram(mergings_ward)
plt.show()


# ### Выбор лучшего разбиения

# In[12]:


mergings_complete = linkage(X, method='complete')
mergings_complete


# In[13]:


dendrogram(mergings_complete)
plt.show()
```

```
# In[14]:


import numpy as np

def update_cluster_centers(X, c):
    centers = np.zeros((4, 2))
    for i in range(1, 5):
        ix = np.where(c == i)
        centers[i - 1, :] = np.mean(X[ix, :], axis=1)
    return centers


# In[15]:


from scipy.cluster.hierarchy import fcluster


# In[16]:


T = fcluster(mergings_complete, 4, criterion='maxclust')
clusters = update_cluster_centers(X, T)
clusters


# In[17]:


plt.scatter(X[:, 0], X[:, 1], c=T)
plt.scatter(clusters[:, 0], clusters[:, 1], c='black')


# ### Вычисление характеристик

# In[18]:


from sklearn.metrics.pairwise import euclidean_distances


# In[19]:


sum_sq_dist = np.zeros(4)
for i in range(1, 5):
    ix = np.where(T == i)
    sum_sq_dist[i - 1] = np.sum(euclidean_distances(*X[ix, :], [clusters[i -
1]]) ** 2)
sum_sq_dist = np.sum(sum_sq_dist) / 4
sum_sq_dist


# In[20]:


sum_avg_intercluster_dist = np.zeros(4)
for i in range(1, 5):
    ix = np.where(T == i)
    sum_avg_intercluster_dist[i - 1] = np.sum(euclidean_distances(*X[ix, :],
[clusters[i - 1]]) ** 2) / len(*X[ix, :])
sum_avg_intercluster_dist = np.sum(sum_avg_intercluster_dist) / 4
```

```
sum_avg_intercluster_dist


# In[21]:


sum_intercluster_dist = np.sum(euclidean_distances(clusters, clusters))
sum_intercluster_dist


# ### Кластеризация выборки методом k-средних

# In[22]:


from sklearn.cluster import KMeans


# In[23]:


models = []
predicted_values = []

for k in range(1, 11):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(X)
    models.append(kmeans)
    predicted_values.append(kmeans.predict(X))


# In[24]:


sum_sq_dist_avg = []
for it, kmean in enumerate(models):
    sum_sq_dist_avg.append(kmean.inertia_ / (it + 1))
sum_sq_dist_avg


# In[25]:


plt.plot(range(1, 11), sum_sq_dist_avg, '-o')


# In[26]:


# Средней суммы средних внутрикластерных расстояний
new_centers = [kmean.cluster_centers_ for kmean in models]

sum_avg_intercluster_dist_avg = []
for k, kmean in enumerate(models):
    intercluster_sum = np.zeros(4)
    for i in range(4):
        ix = np.where(predicted_values[k] == i)
        if len(ix[0]) == 0:
            intercluster_sum[i - 1] = 0
        else:
            intercluster_sum[i - 1] = np.sum(euclidean_distances(*X[ix, :],
[kmean.cluster_centers_[i - 1]]) ** 2) / len(*X[ix, :])
    sum_avg_intercluster_dist_avg.append(np.sum(intercluster_sum) / (k + 1))
sum_avg_intercluster_dist_avg
```

```
# In[27]:


plt.plot(range(1, 11), sum_avg_intercluster_dist_avg, '-o')


# In[28]:


# Средней суммы межкластерных расстояний от количества кластеров
sum_intercluster_dist_avg = []

for k, kmean in enumerate(models):
    value = np.sum(euclidean_distances(kmean.cluster_centers_,
kmean.cluster_centers_))
    sum_intercluster_dist_avg.append(value / (k + 1))
sum_intercluster_dist_avg


# In[29]:


plt.plot(range(1, 11), sum_intercluster_dist_avg, '-o')


# ### Составление сравнительной таблицы

# In[30]:


import pandas as pd


# In[31]:


columns = pd.MultiIndex.from_product([['Иерархический метод', 'Метод k-
средних'],
                                      ['Сумма квадратов расстояний до
центроида', 'Сумма средних внутрикластерных расстояний', 'Сумма межкластерных
расстояний']])
df = pd.DataFrame(columns=columns)
df


# In[32]:


df['Иерархический метод', 'Сумма квадратов расстояний до центроида'] =
[sum_sq_dist for _ in range(len(sum_sq_dist_avg))]
df['Иерархический метод', 'Сумма средних внутрикластерных расстояний'] =
[sum_avg_intercluster_dist for _ in
range(len(sum_avg_intercluster_dist_avg))]
df['Иерархический метод', 'Сумма межкластерных расстояний'] =
[sum_intercluster_dist for _ in range(len(sum_intercluster_dist_avg))]

df['Метод k-средних', 'Сумма квадратов расстояний до центроида'] =
sum_sq_dist_avg
df['Метод k-средних', 'Сумма средних внутрикластерных расстояний'] =
sum_avg_intercluster_dist_avg
df['Метод k-средних', 'Сумма межкластерных расстояний'] =
sum_intercluster_dist_avg
```

```
df
```

```
# In[33]:
```

```
df.to_excel('result.xlsx')
```