Министерство науки и высшего образования Российской Федерации
федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Московский государственный технический университет имени
Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Робототехники и комплексной автоматизации»
КАФЕДРА «Системы автоматизированного проектирования (РК-6)»

ОТЧЕТ О ВЫПОЛНЕНИИ ЛАБОРАТОРНОЙ РАБОТЫ

по дисциплине «Вычислительная математика»

Студент:	Пролыгина Алина Максимовна
Группа:	РК6-51Б
Тип задания:	лабораторная работа
Тема:	Спектральное и сингулярное разложения

Студент _____ Пролыгина А. М.
подпись, дата Фамилия, И.О.

Преподаватель _____ Соколов А. П.
подпись, дата Фамилия, И.О.

Москва, 2021

Содержание

Спектральное и сингулярное разложения **3**
1 Задание . . . 3
2 Цель выполнения лабораторной работы . . . 5
3 Выполненные задачи . . . 5
4 Базовая часть . . . 6
4.1 Реализация функции pca(X) . . . 6
4.2 Подготовка данных и применение к ним функции pca(X) . . . 7
4.3 Зависимость стандартного отклонения от соответствующего номера главной компоненты . . . 8
4.4 Сепарация типов опухолей по первым двум главным компонентам . . . 9
5 Продвинутая часть . . . 9
5.1 Построение лапласиан для трех графов . . . 9
5.2 Доказательство того, что лапласиан графа с n вершинами является положительно полуопределенной матрицей, имеющей n неотрицательных собственных чисел, одно из которых равно нулю. . . . 11
5.3 Анализ спектра и вектора Фидлера графов . . . 12
5.4 Кластеризация графа с помощью сортировки матрицы смежности . . . 17
6 Заключение . . . 19

Спектральное и сингулярное разложения

1 Задание

Спектральное разложение (разложение на собственные числа и вектора) и сингулярное разложение, то есть обобщение первого на прямоугольные матрицы, играют настолько важную роль в прикладной линейной алгебре, что тяжело придумать область, где одновременно используются матрицы и не используются указанные разложения в том или ином контексте. В базовой части лабораторной работы мы рассмотрим метод главных компонент (англ. Principal Component Analysis, PCA), без преувеличения самый популярный метод для понижения размерности данных, основой которого является сингулярное разложение. В продвинутой части мы рассмотрим куда менее очевидное применение разложений, а именно одну из классических задач спектральной теории графов – задачу разделения графа на сильно связанные компоненты (кластеризация на графе).

Требуется (базовая часть).

1. Написать функцию $pca(A)$, принимающую на вход прямоугольную матрицу данных A и возвращающую список главных компонент и список соответствующих стандартных отклонений.

2. Скачать набор данных Breast Cancer Wisconsin Dataset: https:archrk6.bmstu.ru/index.php/f/85484391.
   Указанный датасет хранит данные 569 пациентов с опухолью, которых обследовали на предмет наличия рака молочной железы. В каждом обследовании опухоль была проклассифицирована экспертами как доброкачественная (benign, 357 пациентов) или злокачественная (malignant, 212 пациентов) на основе детального исследования снимков и анализов. Дополнительно на основе снимков был автоматически выявлен и задокументирован ряд характеристик опухолей: радиус, площадь, фрактальная размерность и так далее (всего 30 характеристик). Постановку диагноза можно автоматизировать, если удастся создать алгоритм, классифицирущий опухоли исключительно на основе этих автоматически получаемых характеристик. Указанный файл является таблицей, где отдельная строка соответствует отдельному пациенту. Первый элемент в строке обозначает ID пациента, второй элемент – диагноз (M = malignant, B = benign), и оставшиеся 30 элементов соответствуют характеристикам опухоли (их детальное описание находится в файле https://archrk6.bmstu.ru/index.php/f/854842.

3. Найти главные компоненты указанного набора данных, используя функцию $pca(A)$.

4. Вывести на экран стандартные отклонения, соответсвующие номерам главных компонент.

5. Продемонстрировать, что проекций на первые две главные компоненты достаточно для того, чтобы произвести сепарацию типов опухолей (доброкачественная и

злокачественная) для подавляющего их большинства. Для этого необходимо вывести на проекции каждой из точек на экран, используя scatter plot.

Требуется (продвинутая часть):

1. Построить лапласианы (матрицы Кирхгофа) L для трех графов:

   - полный граф G_1, имеющий 10 узлов;
   - граф G_2, изображенный на рисунке 1;
   - граф G_3, матрица смежности которого хранится в файле https://archrk6.bmstu.ru/index.php/f/854844,

   где лапласианом графа называется матрица $L = D - A$, где A – матрица смежности и D – матрица, на главной диагонали которой расположены степени вершин графа, а остальные элементы равны нулю.

2. Доказать, что лапласиан неориентированного невзвешенного графа с n вершинами является положительно полуопределенной матрицей, имеющей n неотрицательных собственных чисел, одно из которых равно нулю.

3. Найти спектр каждого из указанных графов, т.е. найти собственные числа и вектора их лапласианов. Какие особенности спектра каждого из графов вы можете выделить? Какова их связь с количеством кластеров?

4. Найти количество кластеров в графе $G3$, используя второй собственный вектор-лапласиана. Для демонстрации кластеров выведите на графике исходную матрицу смежности и ее отсортированную версию.

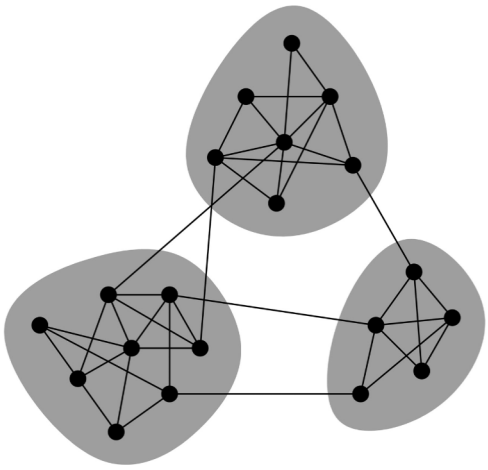

Рис. 1. Граф, содержащий три кластера.

2 Цель выполнения лабораторной работы

Цель выполнения лабораторной работы – рассмотреть метод главных компонент, основанный на сингулярном разложении, и одну из классических задач спектральной теории графов – задачу разделения графа на сильно связанные компоненты (кластеризации на графе).

3 Выполненные задачи

Базовая часть:

1. Разработана функция $pca(A)$, принимающая на вход прямоугольную матрицу данных A и возвращающая список главных компонент и список соответствующих стандартных отклонений

2. Найдены главные компоненты указанного в задании набора данных с использованием функции $pca(A)$.

3. Выведен на экран график зависимости стандартного отклонения от соответствующего номера главной компоненты.

4. Произведена сепарация типов опухолей (доброкачественная и злокачественная) по первым двум главным компонентам.

Продвинутая часть:

5. Построены лапласианы (матрицы Кирхгофа) L для трех графов.

6. Доказано, что лапласиан невзвешенного графа с n вершинами является положительно полуопределенной матрицей, имеющей n неотрицательных собственных чисел, одно из которых равно нулю.

7. Найдены спектры каждого из графов. Выявлены их особенности и связь с количеством кластеров.

8. Найдено количество кластеров в графе $G3$, используя второй собственный вектор лапласиана. Для демонстрации кластеров выведены на графике исходная матрица смежности и ее отсортированная версия.

4 Базовая часть

Все вычисления ниже были реализованы на языке Python, используя библиотеки numpy, matplotlib, pandas и sklearn.preprocessing.

4.1 Реализация функции pca(X)

В базовой части лабораторной работы требуется реализовать метод главных компонент. Для этого необходимо написать функцию $pca(X)$, которая получает на вход прямоугольную матрицу $X \in R^{m \times n}$ и возвращает главные компоненты и стандартные отклонения.

Главные компоненты формируют ортонормальный базис, состоящий из векторов, ассоциированных с наибольшими выборочными дисперсиями.

Формулировка теоремы о главных компонентах:

Главными компонентами матрицы центрированных данных A являются ее сингулярные вектора, при этом j-я главная компонента соответствует j-му сингулярному вектору q_j и стандартному отклонению $\sqrt{\nu}\sigma_j$, где σ_j является j-м сингулярным числом.

Коэффициент ν зависит от выбора оценки ковариации (по умолчанию $\nu = \frac{1}{m-1}$ (несмещенная оценка)).

Следовательно, для нахождения главных компонент необходимо вычислить сингулярные числа и вектора.

Сингулярными числам $\sigma_1, ..., \sigma_r$ матрицы $X \in R^{m \times n}$ называются неотрицательные вещественные числа $\sigma_i = \sqrt{\lambda_i}$, где λ_i – ненулевые собственные числа соответствующей матрицы Грама $K = X^T X$. Ассоциированные собственные вектора матрицы Грама k называют сингулярными векторами.

Следует учитывать свойства сингулярных чисел:

1. Сингулярные числа являются косвенными обобщениями собственных чисел на случай прямоугольных матриц;
2. Матрица Грама K является положительно полуопределенной ($\lambda_i \geq 0$);
3. Сингулярные числа сортируются в убывающую последовательность: $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$;
4. Число r является рангом матриц X и K.

Алгоритм функции pca(X), возвращающей значения главных компонет и стандартных отклонений (листинг 1):

1. Вычисляются собственные числа и вектора матрицы Грама $K = X^T X$ с помощью np.linalg.eigh().
2. Находятся сингулярные числа по формуле $\sigma_i = \sqrt{\lambda_i}$, где λ_i – собственное число.

3. Сингулярные числа сортируются в убывающую последовательность: $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$. Также в соответствии с отсортированными сингулярными числами сортируются сингулярные вектора.

4. По формуле $\sqrt{\nu}\sigma_i$, где $\nu = \frac{1}{m-1}$, вычисляется стандартное отклонение.

Листинг 1. Реализация функции нахождения главных компонент и стандартных отклонений

```
def pca(X):
    w, v = np.linalg.eig(X.T @ X)
    sigmas = np.sqrt(w)
    indexes = np.flip(np.argsort(sigmas))
    sigmas = sigmas[indexes]
    v = v[:, indexes]
    Q = v.T
    variance = np.sqrt(1./(X.shape[0] − 1)) ∗ sigmas
    return Q, variance
```

4.2 Подготовка данных и применение к ним функции pca(X)

Считывание данных в лабораторной работе реализовано с помощью функции read_csv() из библиотеки pandas, ID пациента при этом отбрасываетя, а тип опухоли отделяется от данных по 30 характеристикам и сохраняется в отдельный список. В итоге данные преобразуются в прямоугольную матрицу, в которой 569 строк, каждая из которых соответствует пациенту, и 30 столбцов, которые ассоциируются с характеристиками опухоли.

Для работы с данными необходимо произвести их стандартизацию. Это означает, что данные следует пересчитать по формуле (1), где μ – среднее значение ($\mu = \frac{1}{c}\sum_{i=1}^{c}(x_i)$) и σ – стандартное отклонение ($\sigma = \sqrt{\frac{1}{c}\sum_{i=1}^{c}(x_i - \mu)^2}$).

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

В лабораторной работе стандартизация данных реализована с помощью функции StandardScaler из библиотеки sklearn.preprocessing.

По теореме о главных компонентах для нахождения главных компонет матрицу данных следует центрировать.

Рассмотрение отклонения данных от средних значений, т.е. $a_{ij} = x_{ij} - \bar{x_j}$, дает матрицу центрирования данных A':

$$A = X - e\bar{x} = (E - \frac{1}{m}ee^T)X,$$

где $\bar{x} = \frac{1}{m}e^T X$ – вектор выборочных средних показаний измерений, ee^T – матрица единиц, e – единичный вектор, E - матрица, главная диагональ которой заполнена единицами, а все остальные элементы равны нулю, m – количество строк в матрице.

Центрирование матрицы данных реализовано в функции *get_normalized_data_matrix(X)*, представленной в листинге 2.

Листинг 2. Функция центрирования матрицы

```
1 def get_normalized_data_matrix(X):
2     m = X.shape[0]
3     A = (np.eye(m) − 1./m ∗ np.ones((m, m))) @ X
4     return A
```

Полученная матрица центрированных данных отправляется на вход в функцию pca(X) (листинг 1), которая вычисляет главные компоненты и стандартные отклонения этой матрицы.

4.3 Зависимость стандартного отклонения от соответствующего номера главной компоненты

Метод главных компонент является линейным методом понижения размерности. Для аппроксимации матрицы A другой матрицей меньшего ранга A_k выявляется зависимость стандартного отклонения $\sqrt{\nu}\sigma_i$ от номера главной компоненты (Рис. 2).

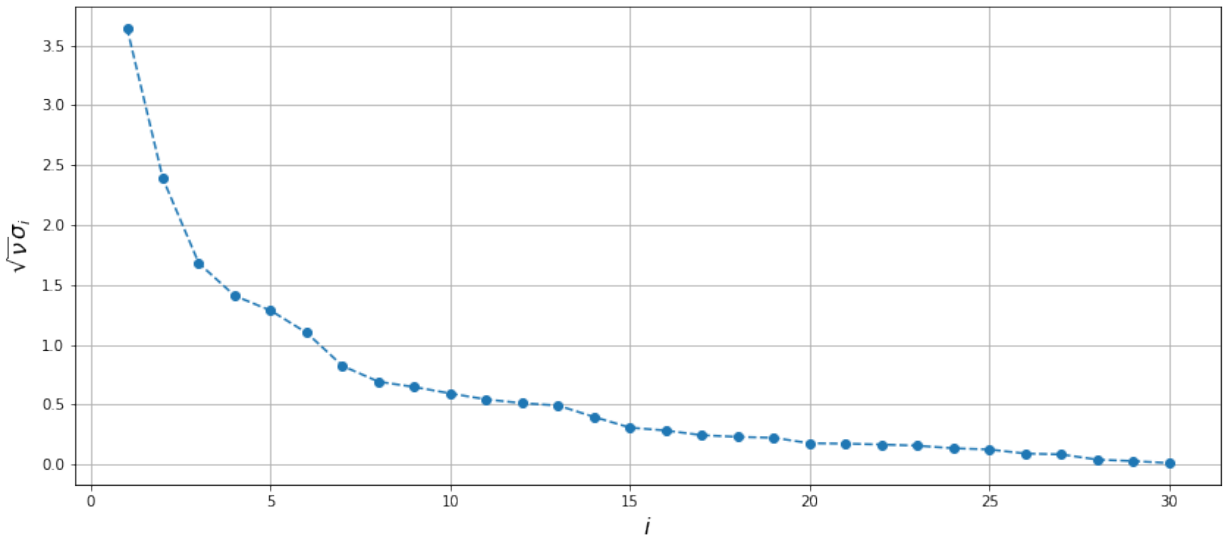

Рис. 2. Зависимость стандартного отклонения от соответствующего номера главной компоненты

По графику на рисунке 2 видно, что наибольшее стандартное отклонение будет у первых двух компонент. Следовательно, можно отбросить остальные характеристики, допустив, что в них содержится меньшая часть информации.

4.4 Сепарация типов опухолей по первым двум главным компонентам

Для сепарации данных необходимо спроецировать их на главные компоненты, т.е. $A_k = A \times Q^T$, где $Q \in R^{2 \times m}$ - матрица, строки которой являются первыми двумя сингулярными векторами матрицы центрированных данных A (возвращается функцией pca(X)).

Результат сепарации представлен на рисунке 3.

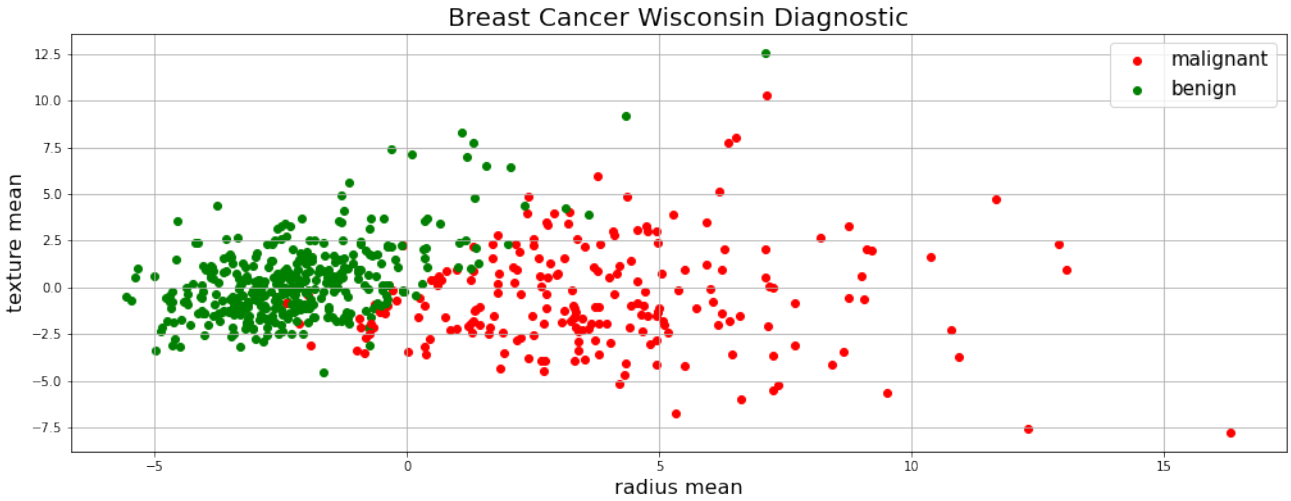

Рис. 3. Сепарация типов опухолей по первым двум характеристикам на доброкачественные (зеленые маркеры) и злокачественные (красные маркеры)

По графику на рисунке 3 можно сделать вывод, что использование двух компонент допустимо для разделения данных диагностики опухолей на две группы по их типу.

5 Продвинутая часть

В продвинутой части лабораторной работы рассматривается задача разделения графа на сильно связанные компоненты (спектральная кластеризация графа).

5.1 Построение лапласиан для трех графов

Лапласианом графа называется матрица $L = D - A$, где A – матрица смежности и D – матрица, на главной диагонали которой расположены степени вершин графа, а остальные элементы равны нулю.

Для нахождения лапласиан была написана функция *laplacian(X)* (листинг 3), которая принимает на вход матрицу смежности графа, а возвращает лапласиан графа.

Листинг 3. Реализация функции нахождения лапласиана графа

```
def laplacian(X):
    D = np.zeros((len(X), len(X)))
    for i in range(len(X)):
        k = 0
        for j in range(len(X)):
            if(X[i][j] == 1):
                k = k + 1
        D[i][i] = k
    L = D - X
    return L
```

В задании требуется найти лапласианы для трех графов:

1. Полный граф G_1, имеющий 10 узлов (Рис. 5). Его матрица смежности является матрицей, главная диагональ которой заполнена нулями, а все остальные элементы равны единице.

2. Граф G_2, имеющий 20 вершин и изображенный на рисунке 1. Матрица смежности A_2 графа G_2 представлена на рисунке 4.

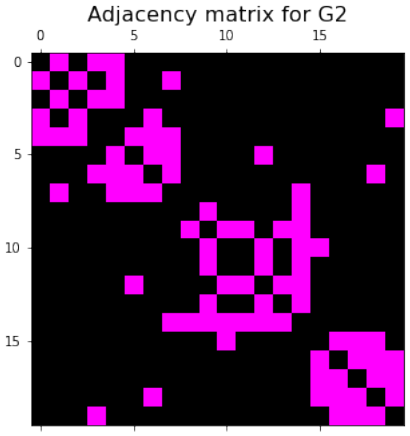

Рис. 4. Визуализация матрицы смежности графа G_2

3. Граф G_3, имеющий 1000 вершин, матрица смежности которого представлена в файле https://archrk6.bmstu.ru/index.php/f/854844.

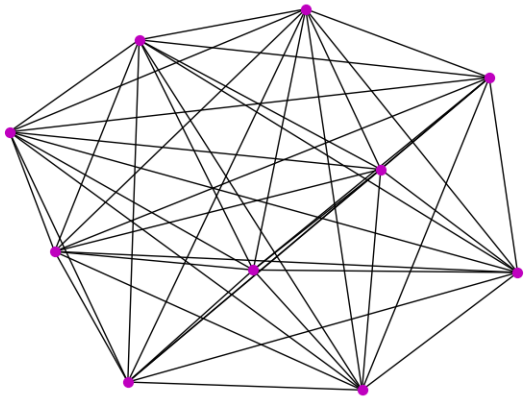

Рис. 5. Визуализация полного графа G_1, имеющего 10 вершин

5.2 Доказательство того, что лапласиан графа с n вершинами является положительно полуопределенной матрицей, имеющей n неотрицательных собственных чисел, одно из которых равно нулю.

Положительно полуопределенная матрица – это матрица, собственные числа которой неотрицательные, т.е. $\lambda_i \geq 0$.

Доказательство того, что лапласиан является положительно полуопределенной матрицей:

Из определения собственного числа: $Lx = \lambda x$, где $x \neq 0$, λ – собственное число. Значит, $\lambda = x^T Lx$.

$$x^T Lx = x^T NN^T x = (N^T x)^T \cdot (N^T x) = \|N^T x\|^2 \geq 0$$

Из $x^T Lx \geq 0$ следует, что L – положительно определенная матрица.

В доказательстве положительной определенности лапласиана использует допущение, что граф произвольно ориентирован. N – ориентированная матрица инцидентности, которая принимает значение 0, если i-тая вершина не принадлежит j-тому ребру графа, значение 1, если i-тая вершина является началом j-того ребра, значение -1, если i-тая вершина является концом j-того ребра. $L = NN^T$.

Доказательство того, что одно из n собственных чисел лапласиана равно 0:

Собственный вектор x является ненулевым решением однородной СЛАУ $(L - \lambda E)x = 0$, где $L \in R^{k \times k}$ – лапласиан, E – матрица, главная диагональ которой заполнена единицами, а все остальные элементы нулевые.

$$B = (L - \lambda E) = \begin{vmatrix} l_{11} - \lambda & l_{12} & \dots & l_{1k} \\ l_{21} & l_{22} - \lambda & \dots & l_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ l_{k1} & l_{k2} & \dots & l_{kk} - \lambda \end{vmatrix} = 0$$

К первой строке прибавляются все остальные строки, учитывая, что сумма элементов строки/столбца равна нулю, получается:

$$B = \begin{vmatrix} -\lambda & -\lambda & \dots & -\lambda \\ l_{21} & l_{22} - \lambda & \dots & l_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ l_{k1} & l_{k2} & \dots & l_{kk-\lambda} \end{vmatrix} = 0.$$

По свойству определителя матрицы выносится λ:

$$B = -\lambda \begin{vmatrix} 1 & 1 & \dots & 1 \\ l_{21} & l_{22} - \lambda & \dots & l_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ l_{k1} & l_{k2} & \dots & l_{kk} - \lambda \end{vmatrix} = 0.$$

Значит, $\lambda = 0$ является решением СЛАУ.

5.3 Анализ спектра и вектора Фидлера графов

Спектром графа называют множество собственных чисел лапласиана графа.

На рисунках 6, 7 и 8 представлены отсортированные собственные значения лапласианов графов G_1, G_2 и G_3, соответственно. По графикам видно, что собственные числа лапласианов L_1, L_2 и L_3 неотрицательные, а первое собственное значение отсортированного спектра каждого из графов равно нулю, что соответствует утверждению, доказанному в предыдущем пункте. Все собственные значения графа G_1 (Рис. 6) за исключением нулевого примерно равны единице, что связано с полносвязностью этого графа.

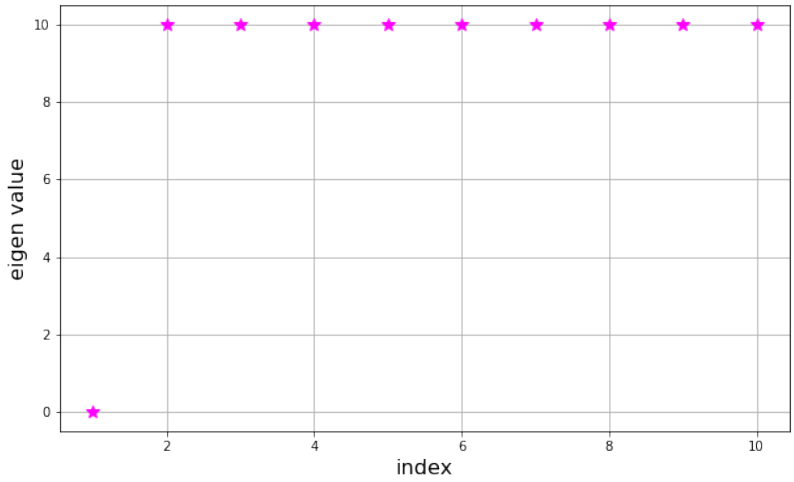

Рис. 6. Спектр графа G_1

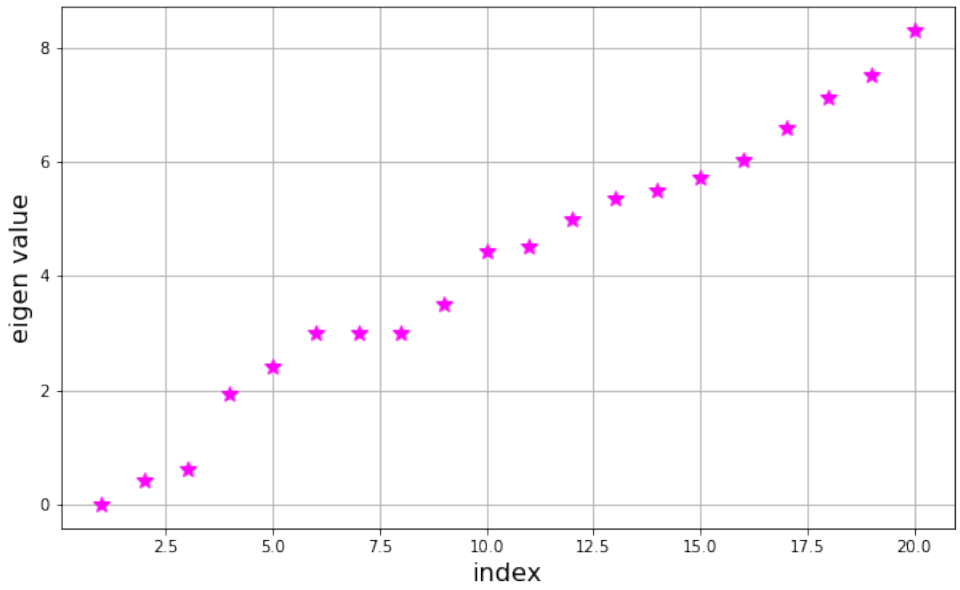

Рис. 7. Спектр графа G_2

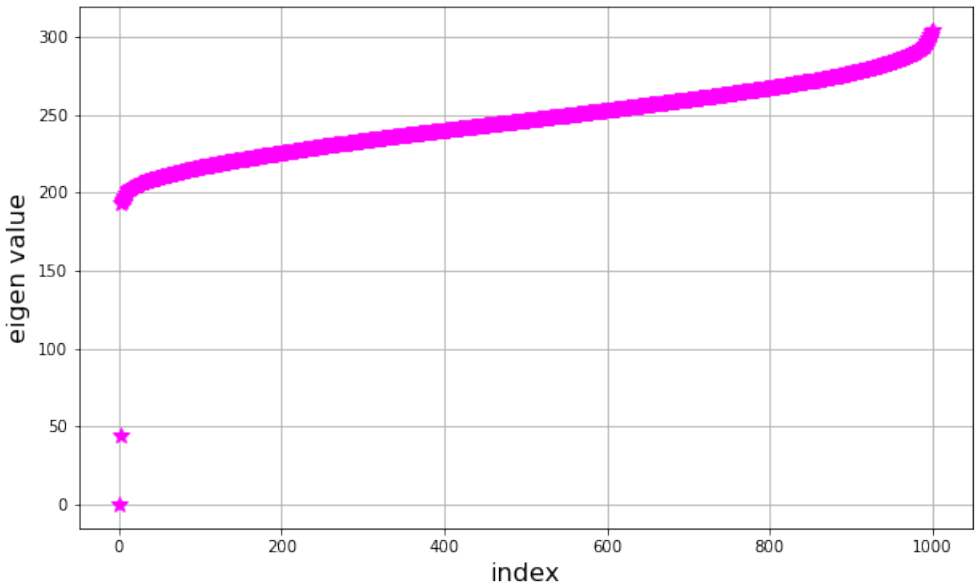

Рис. 8. Спектр графа G_3

Для того, чтобы сделать выводы о связности графов G_2 и G_3, необходимо исследовать вектор Фидлера, который соответствует собственному вектору второго по величине собственного числа лапласиана. Для нахождения вектора Фидлера лапласиана была реализована функция *fiedler_vector()*, представленная в листинге 4.

Листинг 4. Реализация нахождения вектора Фидлера лапласиана

```
def fiedler_vector(X):
    w, v = np.linalg.eig(X)
    f_i = np.argsort(w)[1]
    if(len(X) == 20):
        f_i = f_i + 1
    fiedler = []
    for i in range(len(v)):
        fiedler.append(v[i][f_i])
    return fiedler
```

На рисунках 10, 9 и 11 представлены графики отсортированных векторов Фидлера для графов G_1, G_2 и G_3, соответственно. По графикам легко сделать вывод о количестве кластеров и вершинах, принадлежащих каждому кластеру, для каждого из графов.

Для начала рассматривается вектор Фидлера для графа G_2, т.к. количество его кластеров известно и равно 3. На графике (Рис. 9) видно, что значения отсортированного вектора Фидлера разбиваются на 3 группы. В первой группе – 8 значений, во второй – 7 значений и в третьей – 5 значений. Это соответствует количеству вершин в кластерах графа G_2 на рисунке 1.

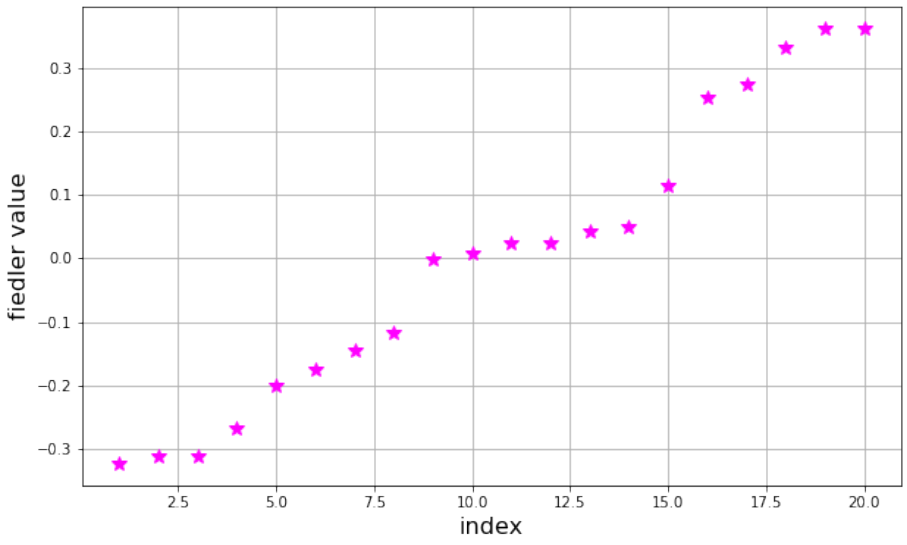

Рис. 9. Вектор Фидлера графа G_2

На графике (Рис. 10) для графа G_1 кластеризации не видно, что является верным, т.к. полносвязный граф G_1 не имеет подграфов.

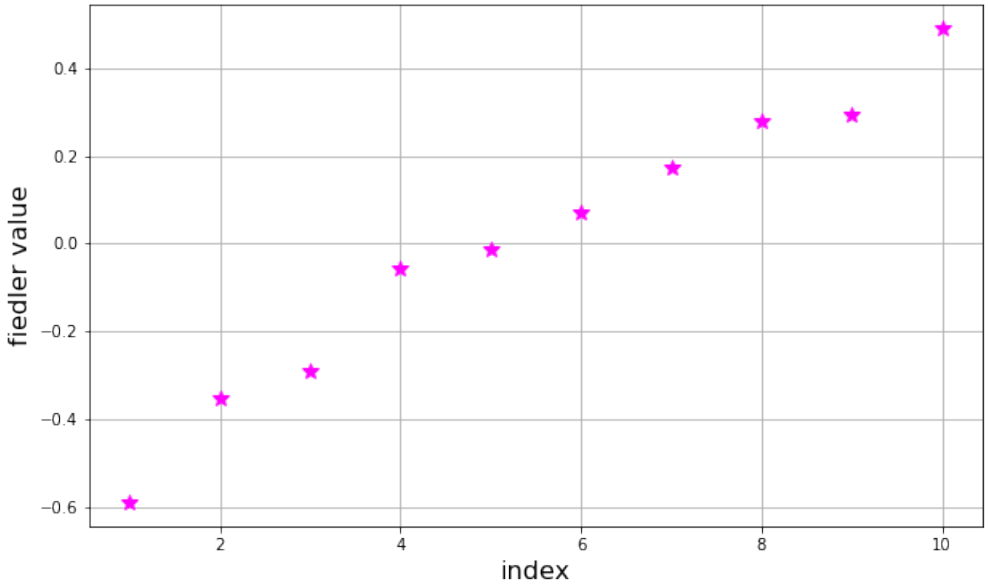

Рис. 10. Вектор Фидлера графа G_1

На графике (Рис. 11) для графа G_3 однозначно различимы 2 кластера. Из-за плотности кластеров и большого количества вершин определить количество вершин, принадлежащее каждому кластеру, по графику невозможно. Можно заметить, что значения вектора Фидлера группируются на отрицательные и положительные. Программным путем было подсчитано, что в векторе Фидлера третьего графа 450 отрицательных и 550 положительных значений. Следовательно, граф G_3 разбивается на 2 кластера по 450 и 550 вершин. Правильность кластеризации графа G_3 подтверждается на рисунке 12 с помощью визуализации из библиотеки networkx.

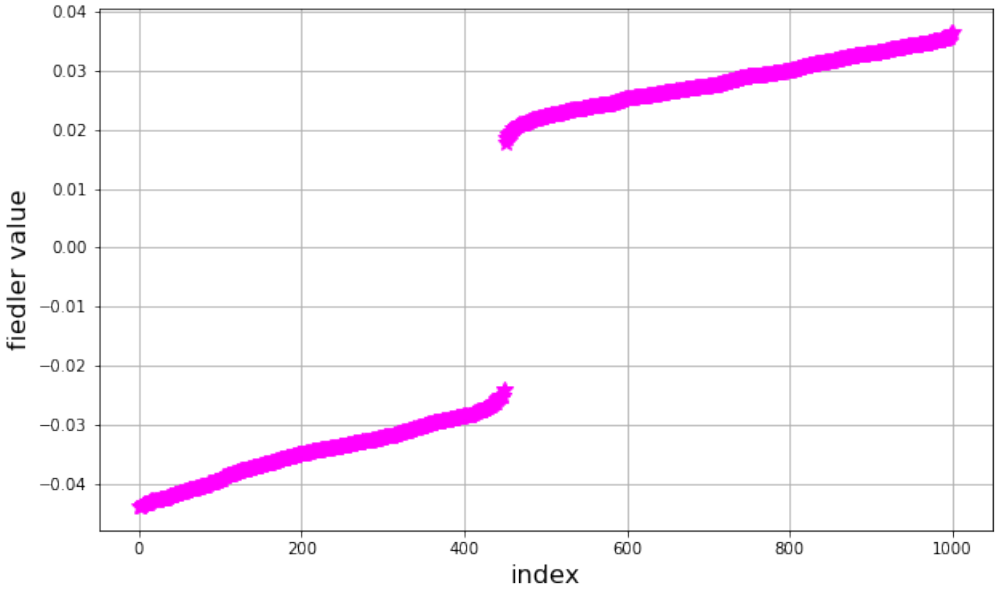

Рис. 11. Вектор Фидлера графа G_3

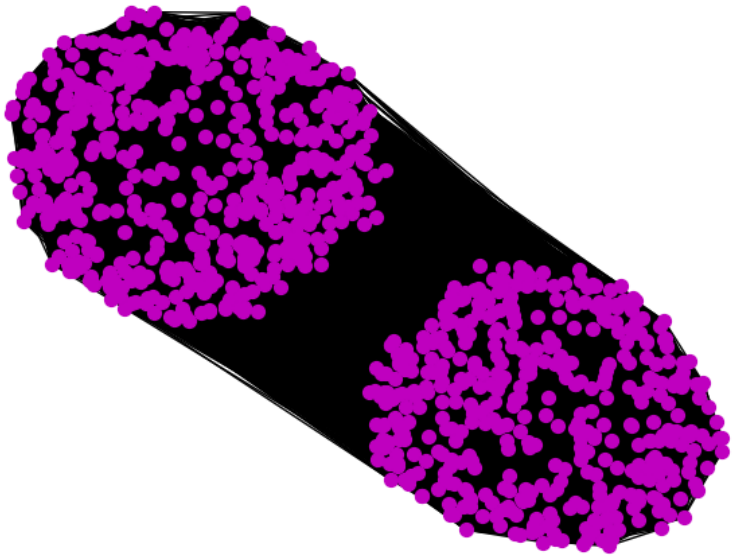

Рис. 12. Визуализация графа G_3

5.4 Кластеризация графа с помощью сортировки матрицы смежности

Для визуализации разбиения графа на матрице смежности следует переставить строки и столбцы матрицы в порядке сортировки вектора Фидлера. Для сортировки матрицы по упорядоченному вектору Фидлера была разработана функция *sorted_matrix(X, fiedler)*, представленная в листинге 5.

Листинг 5. Реализация функции сортировки матрицы смежности по вектору Фидлера

```
def sorted_matrix(X, fiedler):
    s = pd.Series(fiedler).sort_values(ascending=True)
    indexes = s.index.tolist()
    X_sort = np.zeros((len(X), len(X)))
    for i in range(len(X)):
        for j in range(len(X)):
            X_sort[i][j] = X[indexes[i]][indexes[j]]
    return X_sort
```

Проверка функции сортировки проводится на примере графа G_2. На рисунке 13 представлена кластеризация графа G_2, где несложно различить 3 подматрицы, что совпадает с выводами, сделанными в предыдущем пункте, и рисунком 1. Исходная матрица смежности A_2 представлена на рисунке 4.

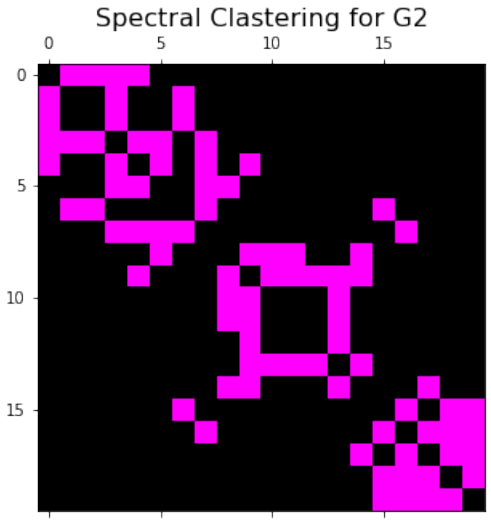

Рис. 13. Кластеризация графа G_2

Визуализация отсортированной матрицы смежности для графа G_3 (Рис. 15) подтверждает выводы, сделанные в предыдущем пункте, что граф G_3 имеет 2 подграфа. Исходная матрица представлена на рисунке 14.

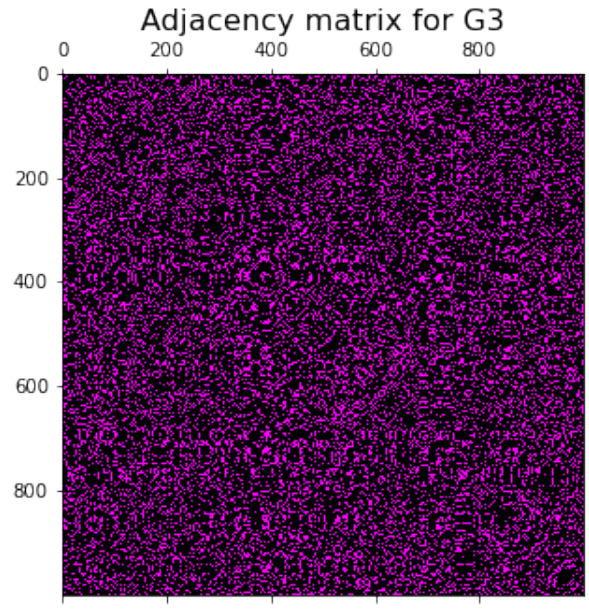

Рис. 14. Матрица смежности графа G_3

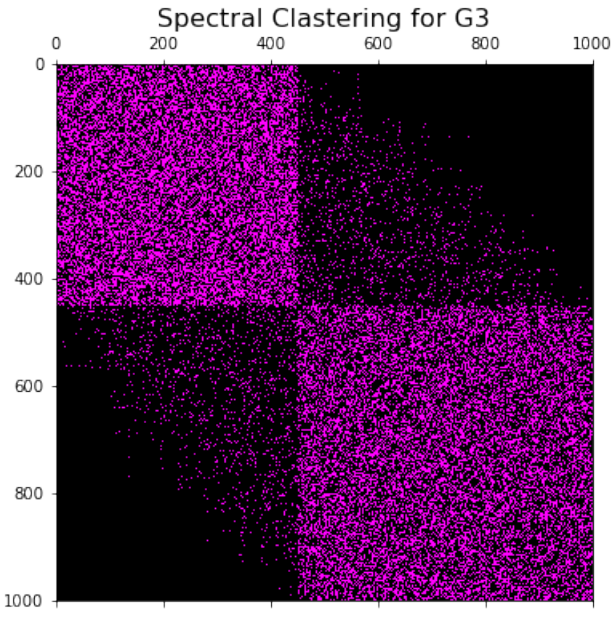

Рис. 15. Кластеризация графа G_3

6 Заключение

1. В базовой части лабораторной работы с помощью метода главных компонент сепарируется набор данных, в котором хранятся результаты обследования пациентов на предмет рака молочной железы, на две группы по первым двум компонентам. В результате проецирования данных на компоненты и визуализации сепарации на графике было выявлено, что первых двух характеристик достаточно для разделения пациентов по типам опухолей.

2. В продвинутой части рассматривается спектральная кластеризация для трех графов с различным количеством вершин и связностью. Доказаны и подтверждены на примерах свойства собственных чисел лапласианов графа. В результате исследования отсортированного вектора Фидлера (вектора, соответствующего второму по величине собственному числу лапласиана) для каждого из графов было выявлено количество кластеров, а также количество вершин, принадлежащих каждому кластеру. Результаты исследования были подтверждены с помощью сортировки исходной матрицы смежности.

Список использованных источников

1. Першин А.Ю. Лекции по курсу «Вычислительная математика». Москва, 2018-2021. С. 140.

2. Документация Python 3.9 [Электронный ресурс] [Офиц. сайт]. 2021. (дата обращения 27.10.2021).

Выходные данные

Пролыгина А. М.. Отчет о выполнении лабораторной работы по дисциплине «Вычислительная математика». [Электронный ресурс] — Москва: 2021. — 19 с. URL: https://sa2systems.ru:88 (система контроля версий кафедры РК6)

Постановка: © *ассистент кафедры РК-6, PhD А.Ю. Першин*
Решение и вёрстка: © *студент группы РК6-51Б, Пролыгина А. М.*

2021, осенний семестр