Введение

Представлены задания краткие вспомогательные сведения для выполнения лабораторных работ в рамках курса «Вычислительная математика». Теоретические материалы представлены отдельно в материалах лекций [1] и включают сведения о методах и подходах, применяемых в вычислительной математике. Представлены постановки задач по следующим тематикам: «Интерполяционные многочлены Лагранжа и Эрмита. Интерполяция сплайнами. Автоматическое дифференцирование.»; «Численное дифференцирование. Численное интегрирование (квадратуры Гаусса, Гаусса–Лобатто). Тригонометрическая аппроксимация. Алгоритм Кули–Тьюки. Дискретное преобразование Фурье. Метод наименьших квадратов. Линейная и нелинейная регрессии. Анализ временных рядов.»; «Прямые и итерационные методы решения СЛАУ. Разложение Холецкого и LU-разложение. Положительно определённые матрицы. Матричные нормы. Модель Лоренца. Вынужденные колебания маятника. Разреженные матрицы.»; «Задача Коши. Методы Рунге–Кутты, Адамса, Адамса–Башфорта, Адамса–Моултона, многошаговые методы численного интегрирования СОДУ. Анализ вычислительной устойчивости». Теоретические материалы представлены отдельно в материалах лекций, в т.ч. представляемых в форме презентаций в процессе обучения.

Инструкция по выполнению [2] лабораторных работ размещена в облачном сервисе кафедры в разделе “70 - Инструкции. Образование”.

4.6 Спектральное и сингулярное разложения (вариант 5)

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).

Задача 25 (спектральное и сингулярное разложения)
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[90].

2. Скачать набор данных Breast Cancer Wisconsin Dataset: https://archrk6.bmstu.ru/index.php/f/854843[91].

– 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 элемент 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 экран проекции каждой из точек на экран, используя scatter plot.

[90] Для нахождения собственных чисел и векторов вы можете использовать функцию `numpy.linalg.eig`.

[91] Данные взяты с сайта https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

6. Опциональное задание №1. Постройте классификатор в полученном пространстве пониженной размерности, используя любой из классических алгоритмов машинного обучения (например, логистическую регрессию или метод опорных векторов) и, при необходимости, кросс-валидацию.

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 2;

– 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[92,93]?

4. Найти количество кластеров в графе G_3, используя второй собственный вектор лапласиана. Для демонстрации кластеров выведите на графике исходную матрицу смежности и ее отсортированную версию[94].

5. Опциональное задание №2. Реализуйте алгоритм DBSCAN и произведите с помощью него кластеризацию графа G_2.

Вопросы и ответы

Вопрос 1

Какой должен быть размер шрифта текстовых подписей, включенных в состав иллюстрации?

Ответ

Шрифт текста на иллюстрациях должен быть сравним со шрифтом подписи к иллюстрации и может быть немногим меньше шрифта основного текста документа.

Комментарий

Разрешение иллюстраций не должно быть ниже 300dpi, что позволит осуществлять некоторое масштабирование без потери качества текстовых подписей.

Вопрос 2

Могут ли использоваться различные шрифты в одном документе (в части размера, курсива, полужирного, типа)?

Ответ

Нет.

Комментарий

[92] В контексте графов кластером называется подграф с большой плотностью связей. Например, граф, изображенный на рисунке 2 имеет три кластера.

[93] Попробуйте взглянуть на собственный вектор, соответствующий второму по величине собственному числу. Его сортировка позволяет выявить не только количество кластеров, но и вершины, принадлежащие конкретному кластеру.

[94] Это можно сделать, например, с помощью функции `matplotlib.pyplot.matshow`

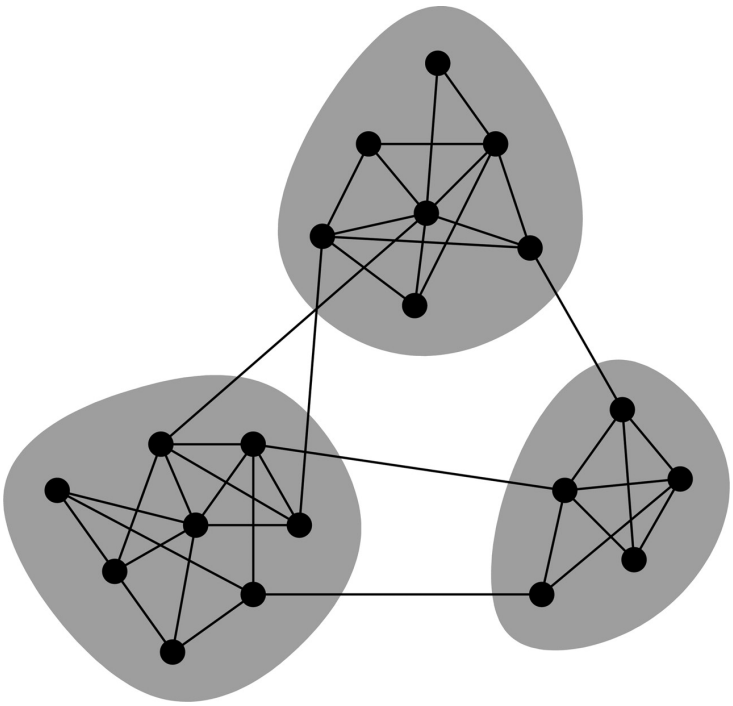

Рис. 2: Граф, содержащий три кластера

Применение различных шрифтов в одном документе для подготовки основного текста недопустимо и является признаком некомпетентности. Каждый шрифт используется для решения специальный задач: выделение заголовков и подзаголовков (увеличенный, полужирный), написание основного текста (обычный), выделение терминов (курсив), подписи к рисункам, таблицам и листингам (уменьшенный, обычный).

Вопрос 3

Какого размера должна быть одна иллюстрация на странице?

Ответ

Субъективно с точки зрения автора: для определения размера одной иллюстрации по ширине текста на странице следует использовать правило золотого сечения.

Комментарий

В дополнение следует отметить, что размер иллюстрации должен быть минимально возможным, но достаточным для представления необходимой информации. Не следует оставлять на иллюстрациях лишние поля и непропроционально большие пустые пространства.

Вопрос 4

Каким форматам следует отдавать предпочтение при подготовке иллюстраций?

Ответ

Векторным (например, EPS) и лишь затем растровым (JPG, PNG) с расширением не ниже 300dpi.

Комментарий

Векторные форматы не зависят от размера области представления, позволяют масштабировать изображение с сохранением качества.

Вопрос 5

Насколько допустима вставка чужих иллюстраций в свои документы?

Ответ

Крайне нежелательна.

Комментарий
Если осуществляется вставка чужих иллюстраций, то это следует делать с обязательной ссылкой на первоисточник. В противном случае такое заимствование может расцениваться максимум как плагиат, и как минимум – некомпетентность.

Список литературы

[1] Першин А.Ю. Лекции по курсу «Вычислительная математика». Москва, 2018-2021. С. 140. URL: https://archrk6.bmstu.ru/index.php/f/810046. (облачный сервис кафедры РК6).

[2] Соколов, А.П., Першин, А.Ю. Инструкция по выполнению лабораторных работ (общая). Москва: Соколов, А.П., Першин, А.Ю., 2018-2021. С. 9. URL: https://arch.rk6.bmstu.ru. (облачный сервис кафедры РК6).