Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего 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	Петраков Станислав Альбертович
Группа	РК6-56Б
Тип задания	лабораторная работа
Тема лабораторной работы	Спектральное и сингулярное разложения

Студент _____ **Петраков С.А.**
подпись, дата *фамилия, и.о.*

Преподаватель _____ **Соколов А.П.**
подпись, дата *фамилия, и.о.*

Москва, 2021 г.

Оглавление

Задание на лабораторную работу 3
Цель выполнения лабораторной работы 4
Выполненные задачи 4
1. Разработана функция для метода главных компонент (Principal Component Analysis) 5
2. Построен график зависимости стандартного отклонения от номера главной компоненты 6
3. Выведены проекции точек на главные направления 6
Заключение 7
Список использованных источников 7

Задание на лабораторную работу

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.

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/854843.
   –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.

Цель выполнения лабораторной работы

Цель выполнения лабораторной работы – написать функции для метода главных компонент (Principal Component Analysis). Построить график зависимости стандартного отклонения от номера главной компоненты, вывести проекции точек на главные направления.

Выполненные задачи

1. Разработана функция для метода главных компонент (Principal Component Analysis)
2. Построен график зависимости стандартного отклонения от номера главной компоненты
3. Выведены проекции точек на главные направления

1. Разработана функция для метода главных компонент (Principal Component Analysis)

Реализована функция *principalComponentAnalysis(A)*, на вход которой подаётся матрица A.

Для начала найдем матрицу центрированных данных, которая находится по формуле:

$$\mathrm{A}_{center} = \left(E - \frac{1}{m} ee^T\right) A,$$

где ee^T – матрица единиц; E – нулевая матрица, у которой на главной диагонали стоят 1; A – данная матрица; m – количество столбцов в матрице A. Для нахождения такой матрицы реализована функция *getNormalizedDataMatrix(A)*.

По теореме о главных компонентах: главными компонентами матрицы центрированных данных A являются её сингулярные вектора, при этом j-ая главная компонента соответствует j-ому сингулярному вектору q_j и стандартному отклонению $\sqrt{\nu}\sigma_j$, где σ_j является j-м сингулярным числом; $\nu = \frac{1}{m-1}$.

Воспользуемся функцией *numpy.linalg.eig* из библиотеки *numpy* при нахождении главных компонент и стандартных отклонений. Функция вычисляет собственные числа и собственные вектора.

То есть главная компонента - собственный вектор, а стандартное отклонение вычисляет как $\sqrt{\frac{1}{m-1}}\sigma_j$, причем сингулярное число σ_j - корень из собственного числа, которое находится с помощью функции *numpy.linalg.eig*.

2. Построен график зависимости стандартного отклонения от номера главной компоненты

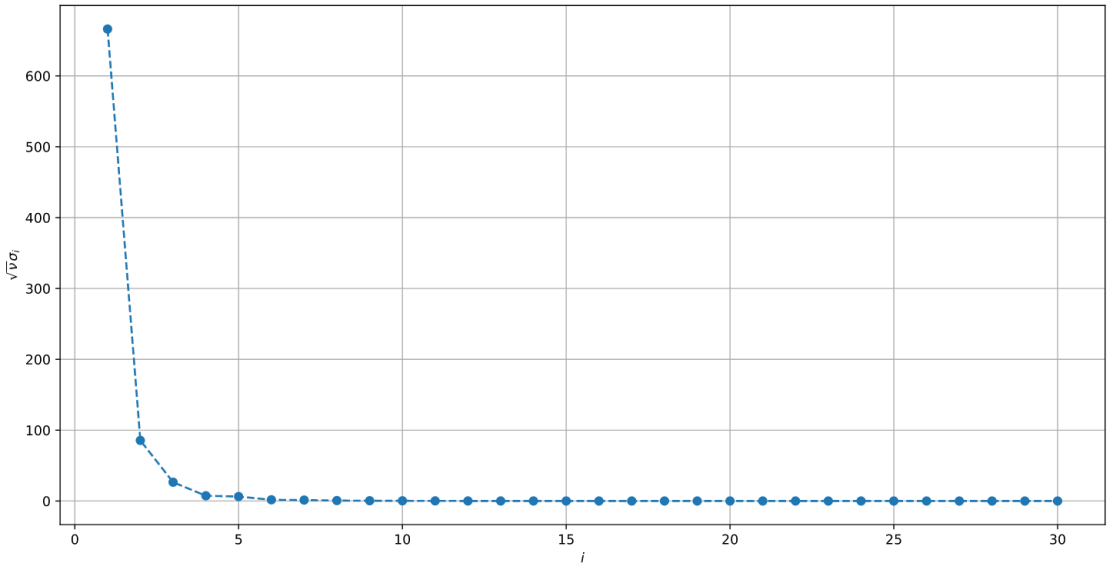

Рис. 1 – Стандартные отклонения, соответствующие номерам главных компонент

По рис. 1 видно, что для первой компоненты соответствует наибольшее выборочное стандартное отклонение, а остальные можно отбросить.

3. Выведены проекции точек на главные направления

По заданию требуется произвести масштабирование признаков. Тогда значения нужно пересчитать по формуле: $x = \frac{x - x_{bar}}{sigma_x}$, где x_{bar} – выборочное среднее значение по данному признаку, $sigma_x$ – выборочное стандартное отклонение.

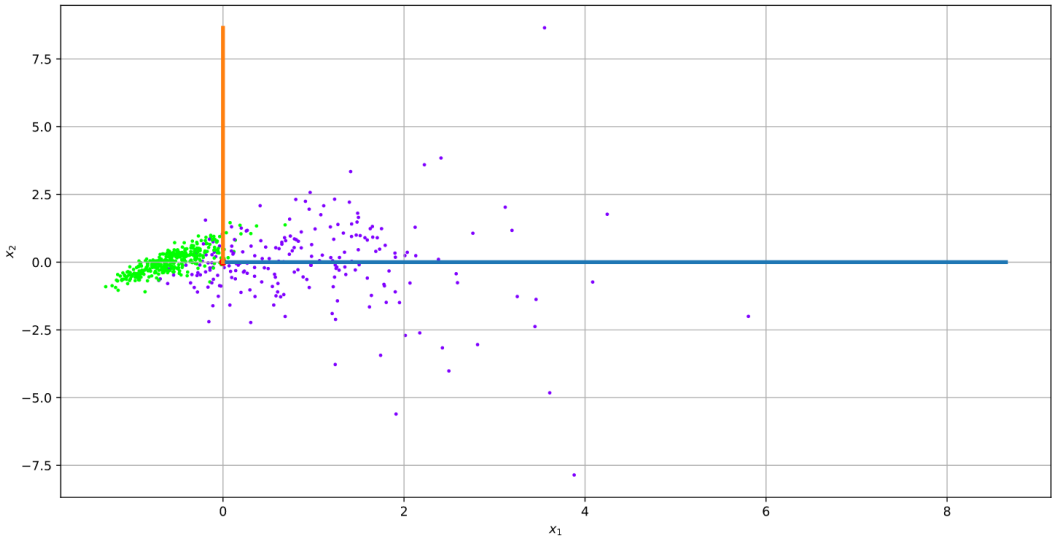

Рис. 2 – Сепарация опухолей; зелёные точки – злокачественные опухоли, фиолетовые – доброкачественные

Заключение

В лабораторной работе был реализован метод главных компонент (Principal Component Analysis), построен график стандартных отклонений, соответствующие номерам главных компонент и проекций точек на главные компоненты.

Список использованных источников

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