УДК 004.85
DOI 10.58649/1694-8033-2023-4(116)-160-166

БУГУБАЕВА Ж.Т.
Ж. Баласагын атындагы КУУ
БУГУБАЕВА Ж.Т.
КНУ им. Ж. Баласагына
BUGUBAEVA ZH.T.
KNU J. Balasagyn

SPIN-код: 2221-4690

СЫЗЫКТУУ АЛГЕБРАНЫН ЫКМАЛАРЫН DATA SCIENCE-ТА КОЛДОНУУ

ПРИМЕНЕНИЕ МЕТОДОВ ЛИНЕЙНОЙ АЛГЕБРЫ В DATA SCIENCE

APPLICATION OF LINEAR ALGEBRA METHODS IN DATA SCIENCE

Кыскача мүнөздөмө: Data Science бизнес-аналитикадан илимий изилдөөгө чейинки тармактарда жаңы түшүнүктөрдү жана чечимдерди камсыз кылуу менен изилдөөнүн негизги багыты болуп калды. Макалада маалымат илиминде маалыматтарды талдоо үчүн сызыктуу алгебра ыкмаларынын ролу жана мааниси талкууланат. Сызыктуу алгебра маалыматтар илиминде негизги ролду ойнойт, маалыматтарды талдоо, иштетүү жана моделдөө үчүн күчтүү куралдар менен камсыз кылат. Макалада сызыктуу алгебранын негизги ыкмаларына жана аларды маалымат илиминде колдонууга сереп берилген. Маалыматтарды манипуляциялоо үчүн колдонулган векторлор, матрицалар, анан алар боюнча операциялар сыяктуу негизги түшүнүктөр каралат. Биз бул түшүнүктөрдү маалыматтарды талдоодо жана иштетүүдө кантип табаарын изилдейбиз. Сызыктуу регрессия методдору, сингулярдык чоңдуктарды ажыратуу жана башка сызыктуу алгебрага негизделген методдор камтылган. Сызыктуу алгебра ыкмаларын колдонуу менен аткарылган кээ бир долбоорлорго жана маселелерге сереп берилет. Векторлор жана матрицалар боюнча операцияларды аткаруу үчүн негизги Python китепканалары да берилген.

Аннотация: Data Science стала ключевой областью исследований, обеспечивающей новые идеи и решения в различных сферах, от бизнес-аналитики до научных исследований. В статье рассматриваются роль и значимость методов линейной алгебры для анализа данных в науке о данных. Линейная алгебра играет ключевую роль в Data Science, предоставляя мощный инструментарий для анализа, обработки и моделирования данных. Статья представляет обзор основных методов линейной алгебры и их применение в Data Science. Рассматриваются такие базовые понятия, как векторы, матрицы и далее операции над ними, используемые для манипуляции данными. Исследуем, как эти концепции находят своё применение в анализе и обработке данных. Рассматриваются методы линейной регрессии, разложение на сингулярные значения и другие методы, основанные на линейной алгебре. Дан обзор некоторых проектов и задач, выполненных с использованием методов линейной алгебры. Также представлены основные библиотеки Python для выполнения операций над векторами и матрицами.

Abstract: Data Science has become a key area of research, providing new insights and solutions in areas ranging from business analytics to scientific research. The article discusses the role and significance of linear algebra methods for data analysis in data science. Linear algebra plays a key role in Data Science, providing powerful tools for analyzing, processing and modeling data. The article provides an overview of the main methods of linear algebra and their application in

Data Science. Basic concepts such as vectors, matrices, and then operations on them used for data manipulation are considered. We explore how these concepts find their application in data analysis and processing. Linear regression methods, singular value decomposition, and other linear algebra-based methods are covered. An overview of some projects and problems carried out using linear algebra methods is given. The main Python libraries for performing operations on vectors and matrices are also presented.

Негизги сөздөр: маалыматтарды талдоо жана иштетүү; Data Science; NumPy, Google Colab; жасалма интеллект.

Ключевые слова: анализ и обработка данных; Data Science; NumPy; Google Colab; искусственный интеллект.

Keywords: data analysis and processing; Data Science; NumPy; Google Colab; artificial intelligence.

Data Science – деятельность, связанная с анализом данных и поиском оптимальных решений на их основе. В основе области лежит наделение смыслом массивов данных, визуализация, сбор идей и принятие решений на основе этих данных.

С помощью науки о данных специалисты изучают данные в разных их проявлениях, находят инсайты и закономерности, моделируют процессы. При этом используют алгоритмы машинного обучения для создания предсказательных моделей. Они в свою очередь применяются уже на новых данных.

Наука о данных – обширная сфера, которая сочетает несколько смежных дисциплин. Это программирование, математика и статистика, бизнес-аналитика и машинное обучение.

Математические знания важны, чтобы уметь анализировать результаты применения алгоритмов обработки данных. Алгебра матриц помогает понять, как большая часть алгоритмов машинного обучения функционирует в потоке данных. Ниже приведены наиболее важные темы для изучения.

1. Основные свойства матрицы и векторов – скалярное произведение, линейное преобразование, транспонирование, сопряжение, ранг, определитель.

2. Внутреннее и внешнее произведение, правило умножения матриц и различные алгоритмы, обратная матрица.

3. Пространственные матрицы – квадратная, единичная, треугольная, разреженная, плотная, симметричная, Эрмитова, антиэрмитова и унитарная матрицы, единичный вектор.

4. Понятие матричного разложения/LU-разложение, метод Гаусса/Гаусса-Жордана, решение систем линейных алгебраических уравнений вида Ax=b.

5. Векторное пространство, базис, оболочка, ортогональность, метод линейных наименьших квадратов.

6. Собственное значение матрицы, собственный вектор, диагонализация, сингулярное разложение (SVD).

Линейная алгебра – это раздел математики, который чрезвычайно полезен в науке о данных и машинном обучении. Она в Data Science и Machine Learning является основополагающей и применяется для решения всех основных задач науки о данных: предварительной обработки и преобразования данных, оценки, создания и настройки моделей, тренировки нейросетей и применения аналитических систем к информации. Все алгоритмы нейронной сети используют ее методы для представления и обработки сетевых

структур и обучающих операций [3]. Кроме того, линейная алгебра применяется в графическом программировании, машинном обучении, обработке сигналов и изображений.

С помощью матриц и векторов многие модели представляют в области обучения с учителем. Это связано с тем, что данные обучения часто имеют вид матрицы, где строки представляют отдельные образцы данных, а столбцы – признаки или характеристики этих образцов. Матричные формы облегчают выражение и понимание алгоритмов обучения, а также позволяют использовать эффективные методы оптимизации. При обучении линейной регрессии [1] задачей является нахождение вектора весов, который умножается на матрицу признаков для предсказания целевой переменной.

Умножение матриц, решение систем уравнений, вычисление градиентов используются при обучении и применении моделей машинного обучения.

Матричное представление данных – мощный инструмент, особенно при работе с несколькими переменными или признаками. Операции сложения, вычитания и умножения в матрице позволяют нам систематически смешивать, изменять и анализировать данные.

Например, умножение матриц представляет возможным комбинировать информацию из разных источников в данных. В машинном обучении [4] оно используется для преобразования признаков и весов модели. Транспонирование матрицы применяют при подготовке данных для анализа, когда признаки представлены в строках, а не в столбцах. Обратные матрицы используют в методе наименьших квадратов для оценки параметров модели. Матричные операции применяют при классификации, регрессии и кластеризации.

Для определения зависимости между признаками используют миноры и детерминанты. В задачах анализа данных может возникнуть необходимость выбрать наиболее информативные признаки из множества доступных.

В аналитике и разработке данных вектор понимают как упорядоченный набор чисел. Этот набор чисел можно представить как координаты, описывающие некоторую точку или движение, либо просто как набор информации. Примерами упорядоченной информации, которую можно описывать векторами, являются, например, координаты ракеты в космосе, биржевые котировки, расположение пикселей в изображении и т.д.

Задача минимизации суммы квадратов расхождений между наблюдаемыми и предсказанными значениями сводится к нахождению такого вектора x, который минимизирует сумму квадратов разностей Ax-b.

Пусть задана матрица признаков X и вектор ответов y. Модуль линейной регрессии может быть представлен уравнением $y = X\beta + \varepsilon$, где β- вектор коэффициентов, ε - вектор ошибок. Задача сводится к нахождению оптимального вектора коэффициентов β. Минимизацию суммы квадратов можно решить с использованием вычисления псевдообратной матрицы. Ее задача заключается в поиске таких параметров модели, которые минимизируют сумму квадратов разностей между фактическими и предсказанными значениями.

Приведем несколько примеров, в которых инструменты линейной алгебры играют важную роль.

Для анализа взаимосвязи между пользователями и товарами необходимо построить рекомендательную систему интернет-магазина, что помогает в предсказании предпочтений пользователей.

Обработку изображений методами PCA используют для снижения размерности признаков; а сингулярное разложение – для выделения наиболее важных компонентов изображения.

Методы линейной регрессии [1] использованы для построения моделей временных рядов. Регрессионная модель полезна для простых временных рядов, или когда есть явные линейности между переменными. В более сложных случаях могут потребоваться более сложные методы. Сингулярное разложение может помочь в анализе зависимостей между различными переменными.

В обработке естественного языка методы векторного представления слов Word2Vec используют линейную алгебру для представления слов в многомерных пространствах.

Методы K-means могут использоваться в пространстве признаков, созданном методами снижения размерности при кластеризации данных [3].

Таким образом, понимание линейной алгебры является ключевым первым шагом на пути к тому, чтобы стать опытным специалистом по Data Science.

Ниже представлен обзор наиболее полезных библиотек Python, которые позволяют выполнять операции линейной алгебры в области Data Science. У каждой из этих библиотек есть свои преимущества – важно понимать, какие операции линейной алгебры необходимы для конкретного проекта и какая библиотека обеспечит нужный баланс функциональности и производительности.

NumPy – универсальный пакет в Python для обработки массивов. Он предоставляет высокопроизводительные объекты многомерных массивов и инструменты для работы с ними. Широко используется в научных и инженерных приложениях благодаря своей производительности и удобству использования. В NumPy размеры называются осями, а число осей – рангом. Он облегчает математические операции над массивами и их векторизацию, а это значительно повышает производительность операций. Данный пакет предоставляет множество функций для выполнения операций линейной алгебры. Вот несколько основных функций, которые предоставляют в этом контексте: с помощью данного пакета можно выполнять основные операции с массивами: добавление, умножение, срез, выравнивание, изменение формы, индексирование массивов; расширенные операции с массивами: стековые массивы, разбиение на секции, широковещательные массивы; работу с DateTime или линейной алгеброй. Функции для вычисления определителя linalg.det(), обратной матрицы linalg.inv(), решения систем линейных алгебраических уравнений linalg.solve(), собственных значений и собственных векторов linalg.eig(), сингулярных разложений linalg.svd() делают NumPy мощным инструментом для работы с линейной алгеброй в Python. Они часто используются в научных вычислениях, статистике, машинном обучении и других областях, где требуются обработка данных и выполнение операций линейной алгебры.

3. **Pandas** – это пакет Python с открытым исходным кодом, который предоставляет высокоэффективные, простые в использовании структуры данных и инструменты анализа для помеченных данных на языке программирования. Он расшифровывается как библиотека анализа данных Python. Pandas – это идеальный инструмент для обработки данных. Он предназначен для быстрой и простой обработки данных, чтения, агрегирования и визуализации.

Pandas берет данные в файле CSV или TSV или базу данных SQL и создает объект Python строками и столбцами, который называется фреймом данных. Фрейм данных очень похож на таблицу в статистическом программном обеспечении, скажем, в Excel или SPSS.

С его помощью можно: выполнять индексирование, манипулирование, переименование, сортировку, объединение фрейма данных; обновить, добавить, удалить столбцы из фрейма данных; восстановить недостающие файлы, обработать недостающие данные или NAN; построить гистограмму или прямоугольную диаграмму [2].

Pandas обеспечивает богатый набор методов для фильтрации, сортировки, группировки, объединения и преобразования данных. Это делает его идеальным инструментом для подготовки данных для анализа.

3. **SciPy** является одним из ключевых пакетов, которые составляют стек SciPy. Теперь есть разница между стеком и библиотекой. Он основывается на объекте массива NumPy и является частью стека, который включает в себя такие инструменты, как Matplotlib, Pandas и SymPy с дополнительными инструментами.

Библиотека содержит модули для эффективных математических процедур, таких как линейная алгебра, интерполяция, оптимизация, интеграция и статистика. Основной функционал этой библиотеки построен на NumPy и его массивах.

SciPy использует массивы в качестве базовой структуры данных. Он имеет различные модули для выполнения общих задач научного программирования, таких как линейная алгебра, интеграция, матанализ, обыкновенные дифференциальные уравнения и обработка сигналов.

С помощью данной библиотеки можно производить математические, научные, инженерные вычисления; процедуры численной интеграции и оптимизации; поиск минимумов и максимумов функций; вычисление интегралов функции; поддержку специальных функций; решение обыкновенных дифференциальных уравнений.

Приведем несколько примеров, в которых инструменты линейной алгебры играют важную роль.

Для анализа взаимосвязи между пользователями и товарами необходимо построить рекомендательную систему интернет-магазина, что помогает в предсказании предпочтений пользователей.

Обработку изображений методами PCA используют для снижения размерности признаков; а сингулярное разложение – для выделения наиболее важных компонентов изображения.

Методы линейной регрессии используются для построения моделей временных рядов. Регрессионная модель полезна для простых временных рядов или, когда есть явные линейности между переменными. В более сложных случаях могут потребоваться более сложные методы. Сингулярное разложение может помочь в анализе зависимостей между различными переменными.

4. **Matplotlib** – это одна из основных и широко используемых библиотек Python для отображения данных, построения графиков и встраивания графиков в приложения. Базовое построение графиков для представления данных очень напоминает MATLAB.

С помощью этой библиотеки можно отображать широкий спектр визуализаций [6]: линейные, точечные, круговые, столбцовые диаграммы и гистограммы; диаграммы с областями; контурные графики; поля векторов; спектрограммы, также можно создавать

истории с визуализированными данными. Эта возможность делает его неотъемлемым инструментом для анализа данных, исследований и презентации результатов. Библиотека имеет обширную документацию и активное сообщество, что облегчает изучение и использование ее возможностей для создания красочных и информативных графиков. Она предоставляет гибкие настройки и широкий спектр возможностей для построения графиков.

5. **TensorFlow** – это библиотека, которая помогает разработчикам создавать крупномасштабные нейронные сети со многими слоями, используя графики потоков данных. Он также облегчает построение моделей глубокого обучения, продвигает современную технологию ML / AI и позволяет легко развёртывать приложения на базе ML [5].

Одним из наиболее развитых веб-сайтов среди всех библиотек является TensorFlow. Такие гиганты, как Google, Coca-Cola, Twitter, Intel, DeepMin – все используют TensorFlow! Он достаточно эффективен, когда дело доходит до классификации, восприятия, понимания, обнаружения, прогнозирования и создания данных. С его помощью можно распознавание голоса / звуки. Используется в автомобильной промышленности; сфере безопасности; UX/UI; телекоммуникации; при анализе настроений – в основном для CRM или CX; текстовых приложениях – обнаружение угроз; Google Translate, Gmail Smart Reply; распознавании лиц – Facebook's Deep Face, Photo Tagging, Smart Unlock; временных рядах – рекомендации от Amazon, Google и Netflix; обнаружении видео – обнаружении движения, обнаружении угроз в реальном времени в играх, в системах безопасности, в аэропортах.

Если говорить о науке о данных в целом, то актуальность этого направления будет только расти. Каждый день мы сталкиваемся с работой специалистов в области этой науки. Так, маркетплейсы формируют главную страницу в зависимости от предпочтений пользователя, что вычисляется по сложным алгоритмам. Рекомендации музыки, фильмов и роликов обрабатываются нейросетями. Все более важной становится предиктивная аналитика, которая применяется для прогнозирования спроса, улучшения и автоматизации продуктов самых разных компаний. Так что и востребованность специалистов по науке о данных растет не то что с каждым годом, а ежедневно.

В заключение: невозможно преувеличить значение линейной алгебры в Data Science. Линейная алгебра, являясь фундаментальным инструментом в науке о данных, играет в ней важную роль, предоставляя мощные инструменты для работы с данными, их обработки и анализа. Это основа многих алгоритмов и методов. Она позволяет эффективно обрабатывать, анализировать и отображать данные, строить и оценивать модели, извлекать информацию из сложных наборов данных и визуализировать. Сложные уравнения также можно решать с помощью методов линейной алгебры, уменьшать размеры и моделировать переменные отношения. Можно находить тенденции, делать точные прогнозы и улучшать модели благодаря тому, что они раскрывают возможности матричных операций, собственных векторов и векторных пространств.

Применение методов линейной алгебры в Data Science обеспечивает исследователей мощными инструментами для извлечения данных, на основе которых принимаются информированные решения.

Список использованной литературы:

1. Грас Дж. Data Science. Наука о данных с нуля. – Санкт-Петербург: Изд-тво «БХВ-Петербург», 2021, 199 с.

2. Кеннеди Б. Основы Python для Data Science. – ДМК «Пресс», 2023, 418 с.

3. Рашка С. Python и машинное обучение. – ДМК «Пресс», 2017, 418 с. – Режим доступа: https://e.lanbook.com/book/100905
4. Силен Д. Основы Data Science и Big Data. – СПб.: ПИТЕР, 2017.
5. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. – Москва: O'Reilly Media, 2017, 392 с.
6. Мюллер Д.П., Массарон Л. Python и наука о данных. – 2-е изд. – СПб., 2020, 241 с.