Вводная часть

Ковальчук Алина Олеговна

12 марта 2021 г.

1 ОБРАБОТКА БОЛЬШИХ ДАННЫХ И ПРОГНОЗНАЯ АНАЛИТИКА

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4}}}} \tag{1}$$

Данное учебное пособие предназначено для студентов, обучающихся по программе высшего образования и направлено на формирование следующих компетенций: - Способность анализировать особенности исходных данных, выбирать адекватные методы решения задач анализа данных; - Способность проводить научные исследования в области методов адаптивного анализа данных; - Способность управлять процессом адаптивного анализа данных.

Пособие построено по модульному принципу. Каждый модуль включает теоретический материал, вопросы для самоконтроля и задания для самостоятельного решения, лабораторные работы.

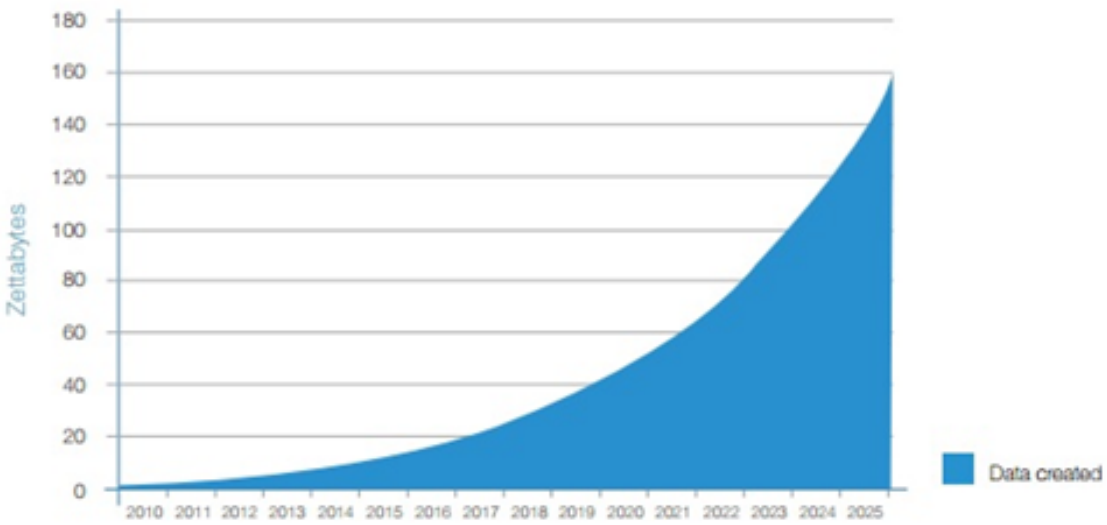
Для успешного освоения материала, представленного в учебном пособии, необходимо владение основами линейной алгебры и математического анализа, а также базовыми навыками программирования на языке Python.

Целями освоения пособия является знакомство студентов с базовыми понятиями и методами анализа данных, примерами их использования в задачах обработки, анализа данных и информационного поиска, а также приобретение навыков аналитика данных и разработчика математических моделей, методов и алгоритмов анализа данных.

1.1 Введение

1.2 Роль анализа данных в современном мире

За последние пару десятилетий привычные нам устройства стали мобильными и подключенными к сети. Многие из нас ежедневно часами сидят в интернете, используя социальные сети, компьютерные игры и поисковые системы. Эти технологические изменения в нашем образе жизни оказали существенное влияние на количество собираемых данных. Подсчитано, что объем данных, собранных за пять тысячелетий с момента изобретения письма до 2003 г., составляет около пяти эксабайт. В наше время такое же количество информации генерируется каждые два дня. По прогнозам IDC количество данных на планете будет как минимум удваиваться каждые два года



Однако сами по себе «данные, как горная порода, – бесполезны без извлекающих золото специалистов и технологий». Сегодня залогом успеха любой компании становится активная работа с имеющейся информацией. Каждый день собирается огромное количество информации о пользователях, их действиях и поведении, информации о развитии бизнеса и маркетинговых компаниях. Современные компании ищут специалистов, способных правильно подготовить и обработать данные, владеющих методами моделирования, анализа и интерпретации результатов их обработки.

Умение понимать и применять числовую аргументацию сегодня востребовано во всех сферах. Анализ данных пронизывает большинство аспектов современной жизни, служит основой

для многих решений в предпринимательской и общественной деятельности, информируют о тенденциях и факторах, которые влияют на нашу жизнь.

Роль методов анализа данных в нашей жизни настолько значительна, что люди, зачастую даже не задумываясь и не осознавая этого, постоянно используют их в повседневной жизни. Принимая решения на работе, делая покупки в магазине и даже знакомясь с другими людьми человек определенным образом анализирует данные, для чего систематизирует и сопоставляет имеющиеся факты, делает необходимые выводы и принимает определенные решения. То есть буквально каждый человек обладает способностями к анализу данных и синтезу информации об окружающем нас мире.

1.3 Анализ больших данных

В современном мире, где информация часто обновляется и поступает из разных источников, специалистам в сфере анализа данных приходится работать с огромными массивами данных. Объем, разнообразие и скорость поступления этих данных создают новые уникальные проблемы для их анализа. Традиционные методы анализа информации не могут угнаться за огромными объемами постоянно растущих и обновляемых данных, что в итоге и открывает дорогу технологиям Big Data.

Технологии Big Data – серия инструментов и методов обработки структурированных и неструктурированных данных огромных объемов и различных форматов. Данные технологии применяются для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста и распределения информации по многочисленным узлам вычислительной сети. Технологии Big Data сформировались еще в конце 2000-х годов в качестве альтернативы традиционным системам управления базами данных и решениям класса Business Intelligence. В настоящее время большинство крупнейших поставщиков информационных технологий для организаций в своих деловых стратегиях используют понятие «большие данные», а основные аналитики рынка информационных технологий посвящают концепции выделенные исследования.

Термин Big Data относится к наборам данных, размер которых превосходит возможности типичных баз данных по хранению, управлению и анализу информации. Введение термина «большие данные» связывают с Клиффордом Линчем – редактором журнала Nature, подготовившему серию работ на эту тему. В настоящее время за развитием технологий Big Data следят множество компаний. В 2011 г. большие данные уже использовались гигантами бизнеса – Hewlett-Packard, IBM, Microsoft. В 2015 г. доля компаний, использующих большие данные, составляла 17% в мире. Сегодня доля таких компаний превышает 50%.

Сегодня технология Big Data активно используется в самых разных отраслях – от банковского до аграрного сектора. Данные становятся таким же важным фактором производства, как трудовые ресурсы и производственные активы. Благодаря аналитике больших данных компании оптимизируют продажи и логистику, лучше узнают клиентов и, как следствие, разрабатывают наиболее подходящие им предложения.

1.4 Характеристики больших данных

Анализ больших данных (Big Data) начинается с их сбора. Информацию получают отовсюду: с наших смартфонов, кредитных карт, программных приложений, автомобилей. Веб-сайты способны передавать огромные объемы данных. Из-за разных форматов и путей возникновения

Big Data отличаются рядом характеристик, определенных в описательной модели больших данных под названием 3V (VVV): Volume (объёмы данных), Velocity (скорость накопления и обработки данных) и Variety (разнообразие источников и типов данных). Набор признаков 3V (VVV) был создан Douglas Laney из компании Meta Group в 2001 году с целью указания на равную значимость управления данными по всем трём аспектам. Рассмотрим приведенные характеристики подробнее: 1. Volume (объём). Огромные объёмы данных, которые организации получают из бизнес-транзакций, интеллектуальных (IoT) устройств, промышленного оборудования, социальных сетей и других источников, нужно где-то хранить. Еще сравнительно недавно это было существенной проблемой, но развитие систем хранения информации облегчило ситуацию и сделало хранение и доступ к информации доступнее. 2. Velocity (скорость). Чаще всего этот пункт относится к скорости поступления данных в реальном времени. В более широком понимании характеристика объясняет необходимость высокоскоростной обработки из-за изменения темпов прироста информации (всплесков активности). 3. Variety (вариативность). Разнообразие больших данных проявляется в их форматах: структурированные данные из клиентских баз, неструктурированные текстовые, видео- и аудиофайлы, а также частично структурированная информация из нескольких источников. Если раньше данные можно было собирать только из электронных таблиц, то сегодня данные поступают как в формате электронных писем, так и голосовых сообщений.

В дальнейшем возникли интерпретации описательной модели Big Data с четырьмя V (добавлялась veracity – достоверность), пятью V (viability – жизнеспособность и value – ценность), и даже семью V (variability – переменчивость и visualization – визуализация). Но компания IDC, например, интерпретирует именно четвёртое V как value (ценность), подчеркивая экономическую целесообразность обработки больших объёмов данных в соответствующих условиях.

1.5 Инструменты анализа данных

В области анализа данных и интерактивных научно-исследовательских расчетов с визуализацией результатов используется довольно большое количество предметно-ориентированных языков программирования и инструментов – как с открытым исходным кодом, так и коммерческих – например, Python, R, Matlab, Stata, SAS и другие.

Рынок компьютерных программ для статистического анализа данных характеризуется высокой конкуренцией, нередки случаи консолидации и поглощений компаний-разработчиков. Перед пользователями различных категорий встает вопрос выбора оптимального программного продукта для поиска верных ответов на существующие вопросы. Очевидно, что оптимальным является вариант, сочетающий в себе необходимые функциональные возможности, высокое качество работы и умеренную цену. При выборе программы для анализа данных следует учитывать следующие параметры: - соответствие характеру решаемых задач; - объем обрабатываемых данных; - требования, предъявляемые к квалификации пользователя (уровень знаний в области статистики); - имеющееся в наличии компьютерное оборудование.

Сравнительно недавнее появление улучшенных библиотек (прежде всего, Pandas – начало разработки в 2008 г.) для Python сделало его серьезным конкурентом в решении задач манипулирования данными. В сочетании с достоинствами Python как универсального языка программирования это делает его отличным выбором для анализа данных, поэтому на сегодняшний день Python считается одним из наиболее востребованных языков в Data Science.

1.6 Инструменты Python для анализа данных

Python широко применяется для анализа данных. Рассмотрим те инструменты, которые предлагает Python на различных этапах решения аналитических задач.

Этап	Инструменты	Шаги алгоритма
Предобработка данных	Python: Pandas, NLTK, Pymysterm	Постановка задачи, Уточнение задачи, Подготовка данных
Исследовательский анализ данных	Python: Pandas, Matplotlib	Постановка задачи, Уточнение задачи, Подготовка данных
Статистический анализ	Python: Pandas, Matplotlib, Numpy	Постановка задачи, Уточнение задачи, Подготовка данных, Прототип решения
Сбор и хранение данных	Python: Pandas, BeautifulSoup	Постановка задачи, Уточнение задачи, Сбор данных
Анализ бизнес-показателей	Python: Pandas, Matplotlib	Прототип решения
Принятие решений в бизнесе на основе данных	Python: Pandas, Matplotlib, Plotly	Прототип решения, Финальное решение и оформление результатов
Визуализация	Python: Pandas, Matplotlib, Plotly, Bokeh, Seaborn	Финальное решение и оформление результатов
Автоматизация	Python: Pandas, Dash	Подготовка данных, Финальное решение и оформление результатов
Прогнозирование	Python: Pandas, Matplotlib, Sklearn	Прототип решения

Из таблицы видно, что наиболее часто применяются следующие библиотеки: Pandas, Matplotlib и NumPy. Именно эти библиотеки и будут наиболее подробно рассмотрены в учебном пособии.

2 Вопросы для самоконтроля

1. Дайте определение понятия «анализ данных».
2. Приведите примеры применения методов анализа данных.
3. Перечислите основные характеристики больших данных.
4. Перечислите этапы решения аналитических задач.

Содержание