Министерство образования и науки РФ ФГАОУ ВО Дальневосточный федеральный университет «ДВФУ»
Школа естественных наук
Кафедра компьютерных систем

Разработка и апробация электронного учебного пособия «Обработка больших данных и прогнозная аналитика»

Диплом на соискание степени бакалавра

Выполнила:
студент группы Б8117-09.03.02
Ковальчук Алина Олеговна

Научный руководитель:
к.ф.-м.н., доцент ККС
Капитан Виталий Юрьевич

Владивосток 2021

Содержание

1 ОБРАБОТКА БОЛЬШИХ ДАННЫХ И ПРОГНОЗНАЯ АНАЛИТИКА

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

На рисунке 1 обозначено среднее значение по времени. 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.

$$\hat{\Phi}[k,l] \quad = \quad \begin{cases} 0 & \text{if } k,l = 0 \\ S_x[k,l] \cdot H_x[k,l] + S_y[k,l] \cdot H_y[k,l] & \text{otherwise} \end{cases} \tag{2}$$

Пособие построено по модульному принципу. Каждый модуль включает теоретический материал, вопросы для самоконтроля и задания для самостоятельного решения, лабораторные работы. На рисунке 2 обозначено среднее значение по скорости.

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

$$\begin{aligned} J_\lambda(x_2, y_2, s_2) = \iint K_\lambda(x_2, y_2) \cdot \left| m_\lambda \left(\frac{x_2 - x_0}{\lambda \cdot s_2}, \frac{y_2 - y_0}{\lambda \cdot s_2} \right) \right|^2 dx_0\, dy_0 = \\ = K_\lambda(x_2, y_2) \otimes \left| m_\lambda \left(\frac{x_2}{\lambda \cdot s_2}, \frac{y_2}{\lambda \cdot s_2} \right) \right|^2 \end{aligned} \tag{3}$$

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. На рисунке 3 обозначено среднее значение по времени. Посмотрим теперь вопросы (Глава 2)

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 [1].

Однако сами по себе «данные, как горная порода, – бесполезны без извлекающих золото специалистов и технологий»(см. Рис. 1). 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 [2].

Умение понимать и применять числовую аргументацию сегодня востребовано во всех сферах [3]. 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 (см. Рис. 2).

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

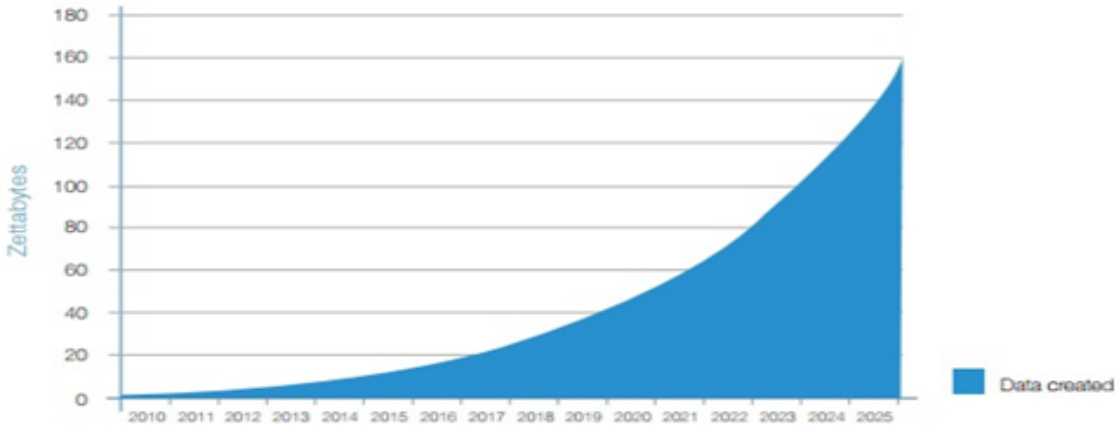

Рис. 1: График

Рис. 2: Обработка данных

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 (см. Рис. 3).

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.

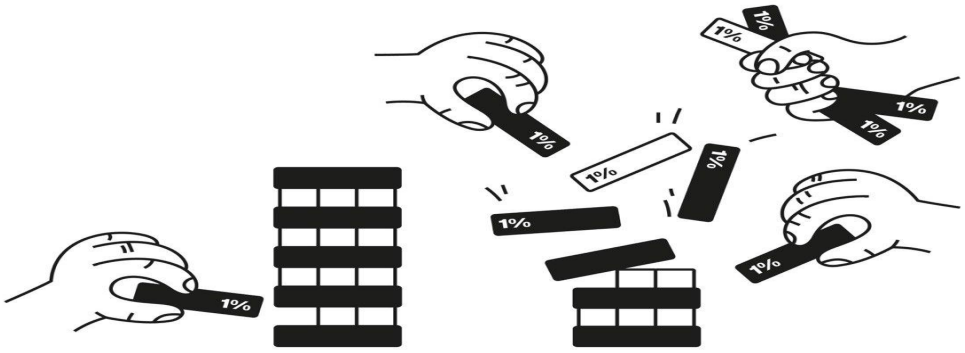

Рис. 3: Красивый рисунок

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. Из таблицы 1 видно, что спутников у нас очень много

Таблица 1: Первые искусственные спутники Земли

ИСЗ	Дата запуска	Масса, кг
Спутник-1	4 октября 1957	83.6
Спутник-2	3 ноября 1957	508.3
Эксплорер-1	1 февраля 1958	21.5

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,

Таблица 2: Инструменты Python

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

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.

Из таблицы 2 видно, что наиболее часто применяются следующие библиотеки: Pandas, Matplotlib и NumPy. Именно эти библиотеки и будут наиболее подробно рассмотрены в учебном пособии. Посмотрим теперь вопросы (Глава 2)

1.7 Графики

Для построения графика в GNUPLOT необходимо найти табличные данные. Сделаем свои данные

Листинг 1: valuegen.py

```
1 import numpy as np
2 import random
3
4 a = range(0,1000,20)
5 b = np.random.randint(0,300,50)
6
7 f2 = open("value.txt",'w')
```

```
for i in range(len(a)):
print(a[i], b[i])
c = str(a[i])+' '+str(b[i])+'\n'
print(c)
f2.write(c)
f2.close()
```

Убедимся, что данные записаны в value.txt. Напишем следующий код для генерации графика в формате .png

Листинг 2: valuerun plt

```
set terminal png
set output 'valueprint.png'
set xlabel "day"
set ylabel "value"
set yrange [0:300]
set xrange [0:1000]
set grid xtics lc rgb '#555555' lw 1 lt 0
set grid ytics lc rgb '#555555' lw 1 lt 0

plot "value.txt" using 1:2 with lines title 'Изменение курса валюты',
    "value.txt" using 1:2:(0.0001) smooth acsplines with lines title
   'Аппроксимация'
```

Запустим скрипт двойным нажатием и получим график:

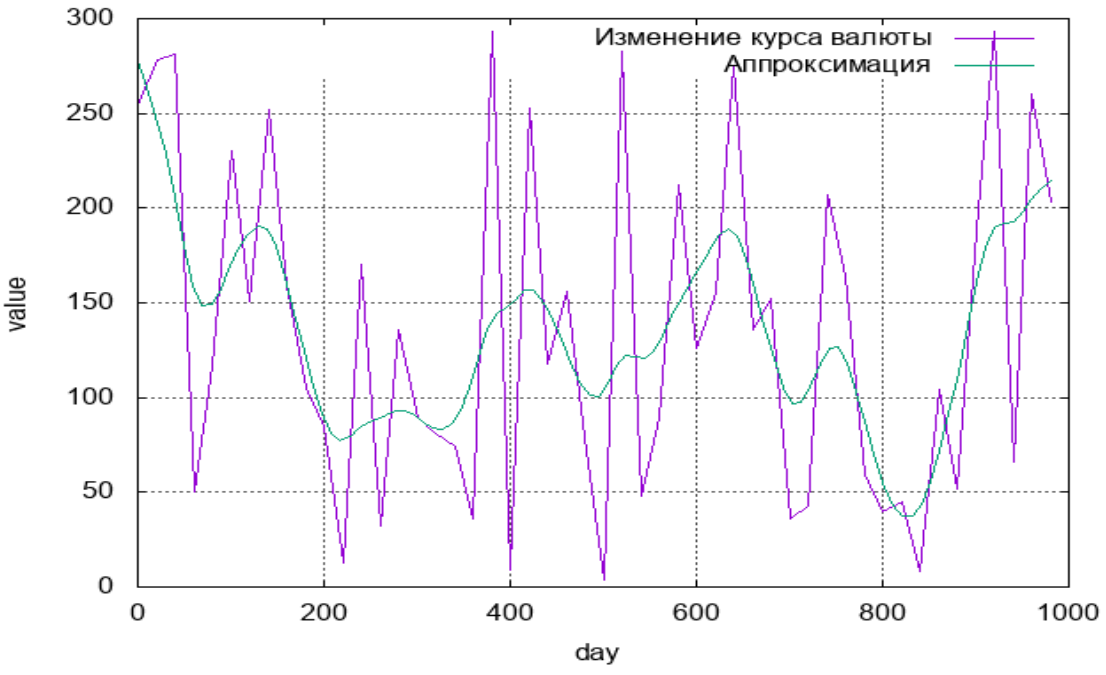

Рис. 4: Прекрасный график для случайных данных

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.

Список литературы

[1] Ronald L Graham, Donald E Knuth, Oren Patashnik, and Stanley Liu. Concrete mathematics: a foundation for computer science. *Computers in Physics*, 3(5):106–107, 1989.

[2] Wil Van Der Aalst. Data science in action. pages 3–23, 2016.

[3] Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambertw function. *Advances in Computational mathematics*, 5(1):329–359, 1996.