CM3005 Data Science

Course Notes

UPDATED APRIL 16, 2021 BY TONI HE

Week 1

1.0 Introduction to the course:

This module will introduce the ***theoretical background*** and ***practical aspect*** of data science.

Key concepts: python programming (NumPy, pandas, MatplotLib), Jupyter notebook, environment, data processing, data visualization, statistics, linear algebra, machine learning.

TOPIC 1: INTRODUCTION TO DATA SCIENCE

This topic will introduce the scope of data science and its impact ins academia and industry.

The main concept in data science will be discussed.

The topic will introduce the Python programming language as the language of choice for many data scientists.

The Jupiter environment for Python development will be introduced and a range of examples will be presented.

TOPIC 2: DATA PROCESSING

This topic will introduce a range of data processing techniques.

The tasks will be demonstrated by using the Python libraries NumPy and pandas.

The topic will present some statistical tasks involving measures of central tendency and measures of spread.

The main concepts of linear algebra will be presented and accompanied with solutions to some practical tasks.

TOPIC 3: INTRODUCTION TO DATA VISUALIZATION

This topic will introduce the main concepts of data visualization.

Different approaches to handling qualitative and quantitative data will be reviewed.

Different types of diagrams and their impact on the intended audience will be discussed and demonstrated.

Real-world examples will be illustrated by using the Python library MatplotLib.

TOPIC 4: STATISTICS

This topic will introduce theoretical foundations of statistics.

The main data types within the context of statistics will be discussed.

The topic will present the theory in which underpins the descriptive and inferential statistics.

The processing of different types of variables will be discussed and demonstrated.

TOPIC 5: MACHINE LEARNING (PART 1)

This topic will introduce the theoretical foundations of machine learning (ML).

The main Python libraries for ML will be discussed and demonstrated.

The concept of model validation and techniques for selecting the best model will be reviewed.

The processing of different types of data using feature engineering will be explored and demonstrated.

TOPIC 6: BASIC TEXT PROCESSING

This topic will introduce fundamental concepts and principles of processing unstructured data.

The main technique use for processing text will be discussed and demonstrated.

The concept of regular expressions and their application to text data will be reviewed.

The practical application of the above techniques to real-world data samples will be explored and demonstrated.

TOPIC 7: NATURAL LANGUAGE PROCESSING

This topic will introduce the fundamental concepts and principles of natural language processing (NLP).

Python libraries for processing natural language data will be discussed and demonstrated.

Techniques for representing word meanings will be reviewed and applied to various NLP tasks.

The practical application of the above techniques within NLP pipelines will be explored and demonstrated.

TOPIC 8: ADVANCED DATA VISUALIZATION

This topic will discuss some advanced visualization techniques.

The topic will discuss large and multidimensional datasets and the challenges accompanying visualization.

Real world examples of multidimensional datasets will be visualized by using:

- Heat Maps
- Parallel coordinates

TOPIC 9: MACHINE LEARNING (PART 2)

This topic will introduce the more advanced techniques in machine learning (ML).

The role of Bayes' theorem and its application to supervised learning will be demonstrated and discussed.

Practical examples of linear regression, support vector machines and decision trees will be reviewed.

Unsupervised learning using kMeans clustering will be explored and demonstrated.

TOPIC 10: CASE STUDIES

This topic will introduce several real-world examples of data science practice.

Topical issues will explore through interviews with practicing data scientists, such as:

- Evaluation techniques applied in industrial settings.
- The skills and competencies they look for in new hires.
- Bias, ethics, and diversity in data science.

1.101 Introduction to Topic 1: The scope of data science

DEFINING DATA SCIENCE:

Data science is a new and emerging subject, still evolving and re-defining itself.

Data science is fundamentally interdisciplinary.

It requires knowledge and skills in three major directions:

Mathematics: more specifically, statistics and linear algebra among others.

Computer Science: machine learning, big data, data visualization etc.

Domain-specific area: the specific area the data is related to.

MAIN PURPOSES OF DATA SCIENCE:

- Recognize sources of data
- Extract insights from data
- Present data in an informative and accessible way
- Enable informed decisions based on data
- Predict future outcomes based on existing data
- Build models of complex systems

WHO USES DATA SCIENCE?

It is widely used in business, science and engineering.

SOME NOTABLE EXAMPLES INCLUDE:

Marketing departments across the corporate world, ie: customer data and metrics.

Social media: user analytics for targeted advertisement.

News outlets: studying targeted audiences

Technology Companies: collect and analyze users' data

IMPACT OF DATA SCIENCE IN PEOPLE'S LIFE:

By using technology, we leave a so called digital footprint

In the age of significant computational and storage capabilities, computer systems have the ability to store and process huge amount of data.

This includes the data we generate as a result of of our activities.

In other words, it is possible to build systems which 'never forget'

In most cases, such systems benefit the customers, since the data allow companies to improve their products and services.

But in other cases, such data might affect people negatively.

A range of data-driven technologies have recently become ubiquitous.

They have the ability to store and process personal information.

SOME PROMINENT EXAMPLES INCLUDE:

- Internet of things
- Wearable devices
- Online services (banking, shopping, services provided by local and national governments)
- Various surveillance technologies.

The utilization of biometrics adds another level of concern, especially from a privacy perspective.

DATA SCIENCE AS A CAREER:

In many cases, data science professionals transition from other fields.

Those include computer science, information technologies, applied mathematics.

Recently, higher education institutions introduced degrees in data science.

SOME POPULAR DATA-DRIVEN CAREER PATHS INCLUDE:

Data Science consulting
Machine Learning Engineer
Big Data Engineer/Architect
Artificial Intelligence Architect

PROMINENT AREAS IN DATA SCIENCE:

Medicine, including drug discovery, medical imaging and diagnosis
Finance and various financial technologies
Social Media
Marketing
Robotics and automation

1.102 The scope of data science

Read Provost, F. and T. Fawcett *Data science for business: what you need to know about data mining and data-analytic thinking*. (Sebastopol, CA: O'Reilly Media, Inc., 2013) Chapter 1 Introduction: Data-Analytic Thinking.

1.103 Nature of data

DATUM NOUN [C] PLURAL: DATA

A journey from Latin to a modern-day buzzword.

Some use ‘data’ as a countable noun, but this is not yet adopted by most major dictionaries.

Can be defined as unit of information or abstraction of a real-world entity.

Used alongside, and sometime interchangeably, with other terms such as: variable, feature, attribute and especially ***information***.

Generally, in academic and scientific contexts the terms ***data*** and ***information*** have different meanings.

In most cases, entities subject to a data science study are described by a number of attributes [1] :

For example:

Entity: Book
- Attributes: Author, Title, Genre, Publisher, Year, Price

Entity: Vehicle
- Attributes: MPG, Cyl, Engine, HP, Weight, 0to60, Year, Origin.

[1] This is the case with structured data, as discussed later in this lecture.

DATA POINTS:

A data point can be considered as a single unit of observation.

The most immediate notion of a data point is a numerical value.

This can be integer, real or complex.

Or binary:

0/1 or True/False.

Computer systems are digital (binary) machines, work on any type of data (text, video, audio) is actually done done on digital (binary) representation of that data.

Programming languages and software systems add an abstraction layer, which allows users to process data in their original type and format.

This allows users to directly apply domain-specific functions on the data.

For example, in audio and image processing and editing.

DATASET

Most generally, a dataset is a collection of data points.

Usually two-dimensional, with rows representing entities (items) and columns representing attributes (features).

Entities, or items in most cases are part of a collection, such as books in a library or items in a shop.

Attributes, or features, describe the entities can be of different types.

ISSUES WITH DATA SETS:

Real world data is not always consistent

The most common issues include:
- missing attributes
- attributes of incorrect type of with incorrect values.

It requires pre-processing (cleaning).

In many cases, this is equivalent to transforming a table into first normal form in the context of a relational database.

STRUCTURE OF DATA

Datasets can be classified based on their structure:

Structured data
- Usually organized into tables
- Easy to search, process study

Unstructured data
- Cannot be organized into consistent structure
- Most real-world data are unstructured

DATA TYPES

The concept of data types is key in statistics and in computer science

The term has mostly overlapping meanings across both areas, although some nuances in the meanings exist.

DATA TYPES IN STATISTICS

Categorical: nominal, ordinal
Numerical: interval, ratio

DATA TYPES IN COMPUTER SCIENCE

Varying depending on the programming paradigm (concurrent, parallel, object oriented) and programming language, some languages share similar data types.

Specific language implementations, different interpreters, compilers and IDEs can offer some variations in data type range, precision and required memory.

General grouping of data:

Primitive (built in)
Composite (records, structures, classes)
Abstract (including data structures)

Ie: Boolean, integer, real (float), character, string, reference (pointer)

Datasets are stored in data structures before being processed:

ie: in Python: lists, dictionaries, arrays (NumPy), DataFrames (pandas), graphs (Networkx) etc.

1.104 Nature of data

Read EMC Education Services and EMC Education Services *Data science and big data analytics: discovering, analyzing, visualizing and presenting data*. (Indianapolis, IN: Wiley, 2015) Chapter 1 Introduction to Big Data Analytics and Chapter 2 Data Analytics Lifecycle.

Week 2

1.201 Python language and ecosystem

Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation.

Python is strongly & dynamically typed, designed to combine the computational capabilities of ANSI C and Shell Script. Python key concepts introduced: import, sum, max, lists, dictionaries.

Python library & associations:
- Pandas: data manipulation
- Matplotlib: plotting numerical data
- Scikit-learn: machine learning
- NumPy: n-dimensional array processing, data manipulation

1.203 Python tutorial

The following tutorial covers fundamentals of Python programming. In data visualisation, we will primarily work with higher level data science libraries, such as pandas. However, it is useful to have a basic grasp of Python, so if you are less confident with your Python programming skills, this is a useful resource to look over and keep at hand.

Python tutorial. Python Software Foundation.

1.204 Jupyter: Open platform for data science

Jupyter Notebook:

Open-source web-based application
Documents containing code, equations, visualizations and text
Code can be modified and executed on the fly
Goldsmith's Jupyter Notebook Server:
https://mscds.doc.gold.ac.uk/jupyter

JupyterLab is browser-based integrated development environment (IDE)

1.205 Jupyter Lab: Getting started

Click on the link below to read an overview of how to get started with JupyterLab:
- JupyterLab Getting started: Overview.

1.206 Jupyter user guide

Click the links to read the following Jupyter user guides:

Jupyter user guide: JupyterLab interface
Jupyter user guide: Text editor
Jupyter user guide: Notebooks
Jupyter user guide: Code consoles.

Week 3

2.001 Introduction to NumPy

2.002 NumPy

Now read:

1. NumPy: Data types
2. NumPy: Array creation

Week 4

Week 5

Week 6

Week 7

Week 8

Week 9

Week 10

Week 11

Week 12

Week 13

Week 14

Week 15

Week 16

Week 17

Week 18

Week 19

Week 20

Week 21

Week 22

Learning Objectives

- describe the concepts taught in this course
- apply the concepts taught in this course to solve problems
- analyse algorithms and data structure in terms of the concepts taught in this course

EXAM: September 6, 2021