MECG11503/MEME19803 Programming for Data Analytics Topic 1: Understanding Basic Programming

Dr Liew How Hui

Jan 2022

Outline

Background

Data science (or data analytics) is about drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference.

- It became popular lately because everyone is trying to immitate FAANG (Facebook, Amazon, Apple, Neflix, Google) and thought that companies became rich because of *mining from data*.
- It is actually old (using the term 'data mining') and has a long history related to the development of statistics.
- It is not a panacea to an organisation.

Background (cont)

- Data Mining: Data + Rules $\Rightarrow$ Output
- Data Science: Data + Output $\Rightarrow$ Rules

	Data Mining	Data Science
Data Sources	Structured Data from RDBMS	Unstructured Data (log files, audio, images, emails, tweets, raw text, documents)
Tools	Data Visualisation, Statistics	Machine Learning
Goals	Knowledge discovery, Decision making	New business value with new functionality (e.g. MS Teams for Online learning, etc.)

Course Outcomes of this Subject

CO1: Demonstrate programming development to turn data into a format suitable for a data science pipeline.

(Topic 1. Group Assignment 20% → A3)

(Topic 4. Assignment 2 (5%) → A3)

CO2: Interpret data using exploratory data analysis. (Topic 2. Practical Quiz 10% → C5)

CO4: Create graphical representations of data.

(Topic 6. Assignment 2 (15%), Practical 2 (15%). By Dr Goh)

CO3: Manipulate data by creating new features, reducing dimensionality, and by handling outliers in the data.

(Topic 3 and Topic 5. Assignment 20%, Practical 3 (15%). By Dr Ivan Yeo)

Outline

Programming Data Structures

Programming Data Structures (stored in memory):

- Basic/Simple Data Structures (Atoms):
  - Boolean data types
  - Integral data types
  - Floating point data types
  - String data types
- Container Data Structures (Chained Molecules): Tuple, List, Set, Dictionary (Hashtable, Associative array)
- Record (or more general, Object-Oriented) Data Structures (Compound Molecules): class, dataclass, etc.

File Data Structures

Real-world data are **compound** (**complex**) **file** data structures (stored in harddisk, SSD or cloud files):

- Word (.doc, .docx), Excel (.xls, .xlsx), Powerpoint (.ppt, .pptx): contains text, images, file attributes, etc.
- PDF documents (.pdf)
- Images (.bmp, .jpg, .png, .webp)
- CSV, HTML, JSON, etc.

They are read and stored as record data structures which can be “broken down” into a collection of organised ‘basic data structures’.

Basic Data Structures (cont)

Basic data type — Boolean (`bool` in Python)

- Values: True, False
- Logical connectives: and, or, not

Application:

- Predicates: They are functions that return Boolean values (and are introduced in discrete maths). E.g. quadraticdiscriminant(a,b,c) := $b^2 - 4ac > 0$
- Predictes or compound predicates are used in if statements, while loops, etc.

Basic Data Structures (cont)

Basic data type — Integer (`int` in Python)

- Values: ..., -3, -2, -1, 0, 1, 2, 3, ... (can be arbitralily large in Python but not Numpy array)
- Arithmetic Operations: +, -, *, /, //, **, %
- Relational Operations: >, >=, ==, ! =, <=, <

Applications

- Representing ordinal data
- Counting
- Indexing in a for loop.

Beware: Python integer is bignum is other system. Numpy has fixed-bit integers int16, int32, int64.

Basic Data Structures (cont)

Basic data type — Floating-point Number (`float` in Python)

- Scientific notation: $\pm$ X.DDDDD $\times 10^{y}$. E.g speed of light `c = 2.99792458e+08` ms^{-1}
- Arithmetic Operations: +, -, *, /, **
- Relational Operations: >, >=, ==, ! =, <=, <
- math library

Applications:

- Approximating continuous numbers in Physics, engineering, financial problems

Basic Data Structures (cont)

Basic data type — String (`str` in Python)

- ”a string”, ’a string’, ”””multiline strings”””
- Operations: +, len(), ”repeat”*5

Applications:

- Unicode text data represention
- “categorical data”, e.g. blood types (A, B, AB, O), etc.

Remark: Byte strings are lower level than strings in Python and will be skipped.

Outline

‘Program Flow’ is sort of like ‘cooking process’ for computer.

- Ingredients: Container data structures or record data structures
- Operations (wash, cut, fry, etc.): functions (they are either available in the Python system or need to be imported from Python libraries)
- Steps: A sequence of steps (programming terminology: statements), a step can be an assignment statement, applying functions, if-else statements, for loops, while loops, etc.

More Data Structures

The basic data sructures (Boolean, integers, floating point numbers and strings) are not good enough for programming. We need more organisations in data structures like:

- Records: A (data) class with named items with multiple items, e.g. CustomerRecord(Name="Ali", Age=18, Address="KL", Phone=12345678)
- Collections / Containers:
  - Tuple: (a, b, c)
  - List: [a1, a2, a3, ...]
  - Set: {a1, a2, a3, ...}
  - Dictionary: {key1:val1, key2:val2, ...}

Record Data Structures

Example: customer information record:

```
from dataclasses import dataclass

@dataclass
class CustomerRecord:
    Name : str
    Age  : int      # or float? Is 21.8 years old OK?
    Address : str
    Phone : int     # or string?

print(CustomerRecord("Ali", 18, "KL", 12345678))
```

Collections / Containers

Tuple: (a,b,c)

Application:

- Simple assignment: a,b,c = 1,2,3
- Encoding error: (result, error)
- Construct array of certain shape: np.zeros((2,2,3))

Containers (cont)

List : super-powerful but slow

Application:

- Store arbitrary items: [”string”, 1, 3.14, True]
- Programming: [], append, len
- Construct Numpy array: E.g. np.array([[0.3,0.5],[-0.2,0.7]])
- Construct trees

Containers (cont)

Set: works like finite set in maths

Application:

- Find $A \cap B$, e.g. A=Customers buying A, B=Customers buying B
- Not really a ‘set’ in the mathematical sense because a set of empty set is not allowed in Python but allowed in maths.

Containers (cont)

Dictionary / Associative Map:

Application:

- Key-value pair: {key : value}, mydict[”word”] = 1
- Basic idea for key-based NoSQL database
- Count items, e.g. mydict[”word”] += 1
- Construct data frame

Functions

In a programming language, a function takes in a value or values and return a value. Note that the value can be basic data types or collection / record data types.

Example 1: Mathematical functions, e.g. sin, cos, exp, etc. from the `math` library.

Example 2: `len` can be used find the length of some lists or sets. E.g. len(set(['a', 'A', 'a', 'b']).union([1,'a',3]))

Example 3: Predicates (mentioned earlier) takes in values and returns a boolean. E.g. def is_positive(x): return $x > 0$

Functions (cont)

User-defined functions are just functions defined by a user following the Python programming syntax. For example, if we want to define a function which calculates sine using the degree as unit, then, we can define our own sine function using the 'sin' from math library to do the real computation:

```
def sine(Degree):
    from math import sin, radians
    return sin(radians(Degree))
```

Note: the function `radians` multiply $\pi/180$ to `Degree` turning it to the radian unit that Python's `sin` can perform numerical calculation.

Program Flow Structures

The various data structures and functions can be orgnised in different programming paradigms to form a program flow structure / architecture:

- Imperative Programming
- Object-Oriented Programming (Complicated for beginners)
- Functional programming (Complicated even for programmers? Welcome by math enthusiast?)

Imperative Programming

It is based on the ‘change of states’ and the following structures:

- if statement : decision
- for loop : going through a list
- while loop : indefinite loop
- break and continue
- defining functions

Examples of Imperative programming languages: C, C++, Python, Go, Fortran, Pascal, MODULA2, MODULA3, ALGO 60, ADA, COBOL, etc.

Imperative Programming (cont)

Example: Simple linear regression:

$$y_i = ax_i + b, \quad i = 1, \ldots, n.$$

The estimate for a and b are as follows (given in SPM and university year 1 statistics):

$$a = (n \sum x_i y_i - \sum x_i \sum y_i)/(n \sum x_i^2 - (\sum x_i)^2)$$
$$b = (\sum x_i^2 \sum y_i - (\sum x_i y_i)(\sum x_i))/(n \sum x_i^2 - (\sum x_i)^2)$$

Imperative Programming (cont)

Implementation using Year 1 Programming Technique in Python:

```
1  import pandas as pd
2  Default = pd.read_excel("Default.xlsx", index_col=0)  # Need openpyxl
3  def convert(x): return 0 if x=="No" else 1
4  Y = Default.default.apply(convert).values
5  X = Default.balance.values
6
7  if X.shape != Y.shape:
8      print(f"Error length(X)={len(X)} != length(Y)={len(Y)}")
9  n = Y.shape[0]; D = range(n)
0  SX = SY = SX2 = SXY = 0.0
1  for i in D: SX  += X[i]
2  for i in D: SY  += Y[i]
3  for i in D: SX2 += X[i]**2
4  for i in D: SXY += X[i]*Y[i]
5  denom = n * SX2 - SX*SX
6  a = (n * SXY - SX*SY)/denom
7  b = (SX2*SY - SXY*SX)/denom
8  print("a=",a,"\nb=",b)
```

Imperative Programming (cont)

Implementation using Year 1 Programming Technique in C++:

```
1  #include <iostream>
2  // https://www.mlpack.org/doc/mlpack-3.0.4/doxygen/formatdoc.html
3  #include <mlpack/core.hpp> // g++ -fopenmp default1a.cpp -lmlpack
4
5  int main() {
6      arma::mat Default;
7      mlpack::data::Load("Default.csv", Default);
8      std::cout << Default.n_rows << " x " << Default.n_cols << "\n";
9      arma::vec Y = arma::conv_to<arma::vec>::from(Default.row(1)); // default
0      arma::vec X = arma::conv_to<arma::vec>::from(Default.row(3)); // balance
1      //X.print();
2
3      int n = Y.n_elem, i;
4      double SX = 0.0, SY = 0.0, SX2 = 0.0, SXY = 0.0;
5      for(i=0; i<n; i++) { SX  += X[i];      }
6      for(i=0; i<n; i++) { SY  += Y[i];      }
7      for(i=0; i<n; i++) { SX2 += X[i]*X[i];   }
8      for(i=0; i<n; i++) { SXY += X[i]*Y[i]; }
9      double denom = n * SX2 - SX*SX;
20     double a = (n * SXY - SX*SY)/denom;
21     double b = (SX2*SY - SXY*SX)/denom;
22     std::cout << "a=" << a << "\nb=" << b << "\n";
23 }
```

Object-Oriented Programming

It is based on classes and objects and it usually includes the imperative programming structures (otherwise many programmers find it difficult to understand a pure object-oriented programming language):

- Classes = Data + Methods (Functions)
- Objects: Instances of Classes
- Instantiate an object: obj = NewClass()
- Sending 'message' to an object using a method: obj.method(parameters)

Examples of object-oriented programming languages: C++, Java, Kotlin, Python, Ruby, Smalltalk, Objective-C, Swift, etc.

Object-Oriented Programming (cont)

Implementation using Year 2 Object-Oriented Programming in Python:

```
1  import pandas as pd
2  Default = pd.read_excel("Default.xlsx", index_col=0)  # Need openpyxl
3  print(Default.head())
4  print(Default.tail())
5  def convert(x): return 0 if x=="No" else 1
6  Y = Default.default.apply(convert).values
7  X = Default.balance.values
8
9  n = Y.shape[0]
0  denom = n*(X**2).sum()-X.sum()**2
1  a = (n * (X*Y).sum() - X.sum()*Y.sum())/denom
2  b = ((X**2).sum()*Y.sum() - (X*Y).sum()*X.sum())/denom
3  print("a=",a,"\n","b=",b)
```

Note: `Default` belongs to the 'pandas.core.frame.DataFrame' class.

Object-Oriented Programming (cont)

High-level object-oriented programming allows simple Data Processing in Python because many things are **wrapped** inside objects:

```
1  import pandas as pd, numpy as np, statsmodels.api as sm
2  Default = pd.read_excel("Default.xlsx", index_col=0)  # Need openpyxl
3  print(Default.dtypes)
4  def convert(x): return 0 if x=="No" else 1
5  Y = Default.default.apply(convert).values
6  X = Default.balance.values
7
8  X = sm.add_constant(X)
9  model = sm.OLS(Y, X)  # Create Ordinary Least Square object
0  linreg = model.fit()  # Use the method .fit() to fit data
1  print(linreg.params)  # print the 'public' values
2  print(linreg.summary())  # Call the summary() method
3
4  import matplotlib.pylab as plt
5  b, a = linreg.params  # Linear Regression: Y = aX + b
6  plt.plot(X, Y, '+')
7  Xrange = np.linspace(X.min(),X.max(), 100)
8  plt.plot(Xrange, a*Xrange+b,'-'); plt.plt.show()
```

Functional Programming

It is based on functions and it usually difficult because it usually does not include the imperative programming structures:

- Data are similar to imperative programming language (basic and record data structures) but are immutable in general.
- List is a fundamental data structure
- Functions works on data structures recursively most of the time

It is so closely related to discrete mathematics that Examples of functional programming languages: Haskell, Ocaml, SML, Scheme, Racket, etc.

Outline

Data Processing Components

- Data Frame (will be introduced in Topic 3 again with advanced features by Dr Ivan Yeo): 'Collections' of Columns
- Basic Statistical Models: Simple numeric statistics in Topic 2. More basic statistics in Topic 3.
- Complex Statistical Models: Predictive Modelling
- Dashboards
- Report Generation
- Powerpoint Presentation

Python Software Environments

- Python Shell + Text Editor
- Spyder IDE
- Jupyter Notebook: Popular but I don't recommend

Practical 1: Reading data of various formats (using software libraries).

1. To familiarise with the program development environment to edit, compile and execute programs.
2. To read data of various formats using Python's Pandas library.

Software Libraries

Python's standard libraries:

- csv, sqlite, lxml, json, ...: Can handle specific structured and semi-structured data.
- struct (for interpreting bytes as packed binary data), codecs (codec registry and base classes): for general (totally unstructured) data.

Python's Extra / Addon libraries:

- pandas (Topic 3, Dr Ivan Yeo): for loading structured (tabular) data
- PIL (pillow image library): for loading images of specific types

Unstructured Data

Examples:

- text (variable length)
- images
- documents: HTML, XML, SGML, Word, PDF
- audios
- videos
- game data

Structured Data

Examples:

- Company CSVs, Excels (.xls, .xlsx, .xlsb, .xlsm, .xltx, .xltm)
- SQL database

Computer Representations:

- Excel Table ???
- SQL Database: Data definition (schema)
- Numpy Array (Topic 2)
- Data Frame: R's data.frame (1991), Python's pd.DataFrame (11 Jan 2008)

Structured Data (cont)

Python libraries to work with structured data:

```
1 import numpy as np
2 import openpyxl # read Excel 2010 files
3 import xlsxwriter # write Excel 2010 files
4 import xlrd, xlwt # read/write Excel .xls files
5 import pandas as pd
6 import matplotlib.pyplot as plt
7 import seaborn as sns
8 import statsmodels.api as sm
```

Structured Data (cont)

Reading CSV using Python's csv library:

```
1  ### https://docs.python.org/3/library/csv.html
2  import csv, numpy as np
3  alist = []
4  # https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength
5  with open('Concrete_Data.csv', newline='\n') as f:
6      csvreader = csv.reader(f, delimiter=',', quotechar='"')
7      for i, row in enumerate(csvreader):
8          # First row is header
9          if i==0:
0              header = []
1              for item in row:
2                  header.append(item)
3              continue
4          floatlist = []
5          for item in row:
6              floatlist.append(float(item))
7          alist.append(floatlist)
8      tbl = np.array(alist)
9  print(header)
0  print(tbl)
```

Structured Data (cont)

Python's panda library (details by Dr Ivan Yeo in Topic 3) read the CSV file much cleaner compare to the use of `csv` library:

```
1 import pandas as pd
2 fn="Concrete_Data.csv"
3 pop_data = pd.read_csv(fn)
4 print(pop_data.head(4))
5 print(pop_data.tail(4))
6 print(pop_data.index)
7 print(pop_data.columns)
8 print(pop_data.dtypes)
```

Structured Data (cont)

HTML with Tables:

```
<html>
<body>
Table 1
<table>
  <tr><th>Firstname</th><th>Lastname</th><th>Age</th><th>Salary</th></tr>
  <tr><td>Jill</td><td>Smith</td><td>35</td><td>3,000--6,000</td></tr>
  <tr><td>Eve</td><td>Jackson</td><td>?</td><td>6,000--9,000</td></tr>
</table>

Table 2
<table>
  <tr><th>Name</th><th colspan="2">Telephone</th></tr>
  <tr><td>Bill Gates</td><td>55577854</td><td>55577855</td></tr>
</table>
</body>
</html>
```

Structured Data (cont)

Python libraries to read HTML Tables:

```
import pandas as pd
tbls = pd.read_html("tableeg.html")   # requires lxml
for i,tbl in enumerate(tbls):
    print((f" Table {i+1} ").center(70, "="))
    print(tbl)
    print()
```

Structured Data (cont)

JSON = JavaScript Object Notation

Sample 1 (by column?):

```
{
    "Firstname" :    ["Jill", "Eve"],
    "Lastname" :     ["Smith", "Jackson"],
    "Age" :          [35, "?"],
    "Salary" :       ["3,000--6,000",
                      "6,000--9,000"]
}
```

Structured Data (cont)

Sample 2 (by column?):

```
[
{
    "Firstname" :    "Jill",
    "Lastname" :     "Smith, II",
    "Age" :          35,
    "Salary" :       "3,000--6,000"
},
{
    "Firstname" :    "Eve",
    "Lastname" :     "Jackson",
    "Age" :          "?",
    "Salary" :       "6,000--9,000"
}
]
```

Structured Data (cont)

Pandas library supports JSON data:

```
import pandas as pd
df1 = pd.read_json("sample1.json")
print(df1)
df2 = pd.read_json("sample2.json")
print(df2)
```

Structured Data (cont)

SQLite database: Use sqlite & SQL language
or
Python sqlite3 library to read sqlite database

```
from pandas.io import sql
import sqlite3
conn = sqlite3.connect('data.db')
query = 'SELECT * FROM tablename'
tbl = sql.read_sql(query, con=conn,
                   parse_dates={'date': '%d/%m/%Y'})
print(tbl.head())
```

Structured Data (cont)

SQLite Database only supports 5 basic data types:

- NULL. The value is a NULL value.
- BLOB. The value is a blob of data, stored exactly as it was input.
- TEXT. The value is a text string, stored using the database encoding (UTF-8, UTF-16BE or UTF-16LE).
- INTEGER. The value is a signed integer, stored in 1, 2, 3, 4, 6, or 8 bytes depending on the magnitude of the value.
- REAL. The value is a floating point value, stored as an 8-byte IEEE floating point number.

Structured Data (cont)

Other formats:

- SPSS (requires special API from IBM),
- SAS (?)
- Advanced / Commercial SQL Database (requires authentication)
- NoSQL Database (Datalog Query? SPARQL?)
- SAP (?)

Structured Data (cont)

Advanced / Commercial SQL Database supports more data types. E.g. Postgresql (see `http://www.postgresqltutorial.com/postgresql-data-types/`) has richer data structures:

- Boolean;
- Character types such as char, varchar, and text;
- Numeric types such as integer and floating-point number;
- Temporal types such as date, time, timestamp, and interval;
- UUID for storing Universally Unique Identifiers;
- Array for storing array strings, numbers, etc.;
- JSON stores JSON data;
- hstore stores key-value pair;
- Special types such as network address and geometric data.

Structured Data (cont)

Numpy array and Panda dataframes:

- dtype / pd.DataFrame().dtypes
  - b boolean
  - i signed integer
  - u unsigned integer
  - f floating-point
  - c complex floating-point
  - m timedelta
  - M datetime
  - O object
  - S (byte-)string
  - U Unicode
  - V void
- Array integer (32 bits, 64 bits) is not the same as Python integer
- Floating point numbers 32 bits and 64 bits

Structured Data (cont)

```
1  df = pd.DataFrame({
2     'a' : [1,2,3],
3     'b' : [4,5,6],
4     'c' : [7,8,9],
5  })   # default index = 0,1,2,...
6
7  df = df.rename(columns={'a': 'A', 'b' : 'B'})
8  df.dtypes
9  # Changing the data type
0  df['B'] = df['B'].astype("str")
```

Outline

Group Assignment (20%)

Introduction to the Group Assignment:

- Two-member/three-member group assignment.
- Reading original data using Python and
- No preprocessing of data using Excel or any other tools allowed!

Outcome of the Assignment: Demonstrate programming development to turn data into a format suitable for a data science pipeline.

Group Assignment (20%) cont.

The report need to contain

- Distribution of tasks and collaboration to achieve the goal (prepare a good report)
- Background analysis of features and targets in each data
- Using Python to read the data array / table
- Using Python to summarise the statistics of the data
- Explain what sort of business would be associated with the data and the sort of pipeline(s) that may be relevant to the data.