Data Analytics

ECS784P

DR. ANTHONY CONSTANTINOU

&

BHUSAN CHETTRI

TIME-TABLE

Lectures/Tutorials:

- ❑ Tuesday 9:00 – 11:00 AM.
- ❑ Venue : Bancroft: 1.13

Lab classes:

- ❑ Tuesday 14:00 – 16:00 PM.
- ❑ Venue : ITL 2nd Floor
- ❑ First lab on Tuesday 16 January.

Staff:

- ❑ Lectures : Bhusan Chettri and Anthony Constantinou.
- ❑ Labs : Nikesh Bajaj, Lingyun Zhao, Gokhan Solak, Muhammad Irfan and Mohammad Malekzadeh.

Anthony Constantinou
Lecturer

Bhusan Chettri
Associate Lecturer

M. Irfan

Gokhan

Lingyun

Nikesh

M. Mohammad

Topics covered on the module

- ❑ Introduction to Data Analysis
- ❑ Basic Python programming
- ❑ The numerical python, Numpy library
- ❑ Pandas data structures and functions
- ❑ Interacting with files, the web, spreadsheets and databases
- ❑ Advanced data manipulation
- ❑ Data visualisation
- ❑ Scikit-Learn for Machine Learning
- ❑ Statistical approaches to Machine Learning
- ❑ Case studies of Data Analytics projects

Module is grouped into 2 parts

Part-I : Machine Learning with Python – Bhusan Chettri

- ❑ Comprises of python and its core libraries such as Numpy, Scipy, Pandas and Matplotlib.
- ❑ Explore pipeline for design and implementation of data analysis operations.
- ❑ A coursework that enables students to utilize the skills acquired in the theory during the course. Carries 30% of the overall mark.
- ❑ Week 1, 2, 3, 5,6 and 7

Part-II : Probabilistic/Bayesian Machine learning – Anthony Constantinou

- ❑ Comprises of statistical approaches for Machine learning such as Bayesian networks.
- ❑ Week 4, 8,9,10,11 and 12

Learning Outcome

At the end of the course students would be able to apply the learned skillset within python framework for design and analysis of a data analytic system framework that comprises of several key steps such as

- ❑ Data cleaning and transformation
- ❑ Building models for decision making and prediction using Machine learning algorithms.
- ❑ Analysing the results produced by the system.
- ❑ Producing results in a form of plots and diagrams.

Approach and module content

- ❑ The approach used is predominantly technical, from a Computer Science perspective. Though it could be taught from a range of approaches, from very theoretical/mathematical through to very business-oriented.

- ❑ Python as the programming language. We will not go deep into Python, just sufficient to enable you to make effective use of the mathematical, data analysis and machine learning libraries available in the python framework.

- ❑ The module also takes a practical approach, with the emphasis on undertaking practical data analysis tasks.

- ❑ The materials supplied should provide you with the ability to go as deep into any specific topic as you require.

Assessment

Coursework:

- ❑ A data analysis project in a group of 3 (max).

Written exam:

- ❑ Standard format, 4 questions.

Data ?

- A world of data – “Big Data”
- Data is produced and collected on a massive scale world-wide.

Some sources of data ?

Data ?

- A world of data – “Big Data”
- Data is produced and collected on a massive scale world-wide.

Some sources of data:

- ✓ Retail and wholesale transactions.
- ✓ Sensor data
- ✓ Video surveillance
- ✓ Population Census
- ✓ Social media and blogging (Facebook, YouTube etc)

NOTE !!! Data are not in a structured form

Structured data

We will often be concerned with “Structured data”. A deliberately vague term that encompasses many different forms of data, such as

- ✓ Multidimensional arrays.
- ✓ Tabular of spreadsheet-like data in which each column may be a different type (string, numeric, date, etc). This includes most kinds of data commonly stored in relational databases or comma separated text files (csv).
- ✓ Multiple tables of data interrelated by key columns (what would be a primary or foreign key for a SQL user).
- ✓ Evenly or unevenly spaced time series data.

Unstructured data

- ✓ Even though it may not always be obvious, a large percentage of data sets can be transformed into a structured form that is more suitable for analysis and modelling.

- ✓ If not, it may be possible to extract features from a data set into a structured form.

- ✓ As an example, a collection of news articles could be processed into a word frequency table which could then be used to perform sentiment analysis.

Semi structured data ..

- ✓ The growth of the web has led to a lot of “Semi-structured data” : Web pages, documents, multimedia data etc.
- ✓ Such data has some structure; Eg: chapters, sections, paragraphs.
- ✓ But these units are variable in size, contains white spaces and are irregular.
- ✓ This, in part, has spawned the NOSQL database movement. NOSQL databases are then a crucial sources of data for analysis.

Sources of data

We will be looking at connecting to the following types of data sources:

- ❑ CSV (comma separated values) files
- ❑ Excel spreadsheets
- ❑ Relational databases (MySQL)
- ❑ NOSQL databases (MONGODB)
- ❑ Web data/documents

Data to Information
(Data Analytics)

- Having lots of data is not immediately useful, it needs to be aggregated, examined in order to make sense of the endless stream of bytes.
- Data Analytics is the process of transforming raw data into useful and usable information.
- It involves extracting information that is not easily deducible but that, when understood, increase our understanding and often leads to the possibility of performing actions to improve a given situation.

Processes involved in Data Analytics - Overview

a) Develop a mathematical or logical model capable of describing system responses under different levels of precision.

b) Use the model to predict the development of the system or its responses to certain inputs.

c) A measure of the effectiveness of the model is its goodness of predictive power. This depends both on the quality of the modelling techniques employed and on the ability to choose a good dataset on which to build the entire data analysis.

d) Choose appropriate methods of data representation that will expose meaningful information resulting from the chosen inputs and their processing by the model. Good representations of the information will expose information that will otherwise remain hidden.

Processes involved in Data Analytics - Overview

d) Test the data model with another set of data for which we know the system response. This will provide an error calculation and a knowledge of the validity of the model and its operating limits. This facilitates comparison with other models for accuracy and efficiency.

e) Deploy the results of the data analysis. This involves the implementation of the results produced by the data analysis, namely, the implementation of the decisions to be taken based on the predictions generated by the model and the risks that are inherent to this implementation.

Contributing Disciplines

Data Analysis is very multi-disciplinary. Projects vary from small to large, but all require some knowledge of the following disciplines:

Computer Science

- Programming: Python, C++, SQL
- File and database formats: XML/HTML, JSON, XLS and CSV

Mathematics and Statistics

- Bayesian methods, regression, clustering

Machine Learning:

- Knowledge of the specific data domain (Biology, Physics, Medicine etc)

What is the role of domain expertise ?

Types of data

- ❑ Categorical: can be grouped or categorised.
- ❑ Nominal: categorical unordered data.
- ❑ Ordinal: categorical ordered data.
- ❑ Interval: categorical ordered data where the size of the space between categories is the same.
- ❑ Numerical : integer or floating point numbers
- ❑ Discrete : which can take only set of values.
- ❑ Continuous : values which can take any numeric value often in a specific range.

Stages /Phases in Data Analysis

- ❑ Problem definition
- ❑ Data source identification, selection and extraction
- ❑ Data cleaning and transformation
- ❑ Data exploration
- ❑ Choosing modelling approach
- ❑ Model development
- ❑ Model validation/test
- ❑ Visualization and interpretation of results
- ❑ Deployment of the solution

Data cleaning and transformation

Not to be under estimated, these processes often consume a significant proportion of the project's resources.

- ❑ Cleaning : dealing with the missing, wrong or uncertain values
- ❑ Transformation: converting to a format or formats amenable for analysis and comparison.

Approaches to initial data exploration

- Data visualization
- Summarizing data
- Grouping data
- Exploration of relationships between the various attributes
- Identification of patterns and trends
- Construction of regression models
- Construction of classification models

Model development

Models can be classified according to the types of result that they produce:

- ❑ Classification models : categorical results
- ❑ Regression models : these models produce numeric results
- ❑ Clustering models : descriptive results produced by these models

Model Validation

- ✓ Typically done by comparing data produced by the model with data produced by the system under study.

- ✓ By using different test data sets, we can estimate the limits of validity of the generated model.

- ✓ The model may only be valid within a certain data range, or confidence level may vary depending on input data values. This process may include comparative evaluation of a number of models.

- ✓ We often split the data into three sub-sets. Training, validation and test sets. We use training data to train our models, validation data for validating (selecting hyper-parameters) and then test set for final evaluation.

Deployment

The data analyst produces a report describing and discussing the results of the analysis. This report must be understandable to management or the client commissioning the project.

The data analysts report will normally discuss the following issues in detail:

- Results of the analysis
- Possible actions based on the results
- Risk analysis
- Measuring the business impact

Deployment ..

- ✓ When the results of the project include the generation of predictive models, these models can be deployed as a stand-alone application or can be integrated within other software.

Deployment ..

Organisational deployment comprises putting into practice the results of the data analysis. This may take a wide range of different forms depending on the organisation:

- Publishing results.
- Developing or withdrawing information systems.
- Changing some aspect of organisational strategy: research, marketing, product development.
- Changing internal or external processes.

Quantitative and Qualitative Analysis

Quantitative analysis:

- ❑ Involves numeric/categorical data.
- ❑ Enables the development of mathematical models.
- ❑ Supports the drawing of objective conclusions.

Qualitative analysis:

- ❖ May include written textual, video or audio data.
- ❖ Conclusions may include subjective interpretations.
- ❖ Enables the exploration of more complex systems not amenable to a strictly mathematical approach, e.g. social phenomena or complex structures which are not easily measurable.

Some sources of open data

- Datahub (http://datahub.io/dataset)
- World Health Organization (http://www.who.int/research/en/)
- Data.gov (http://data.gov)
- European Union Open Data Portal (http://open-data.europa.eu/en/data/)
- Amazon Web Service public datasets (http://aws.amazon.com/datasets)
- Facebook Graph (http://developers.facebook.com/docs/graph-api)
- Healthdata.gov (http://www.healthdata.gov)
- Google Trends (http://www.google.com/trends/explore)
- Google Finance (https://www.google.com/finance)
- Google Books Ngrams (http://storage.googleapis.com/books/ngrams/books/datasetsv2.html)
- Machine Learning Repository (http://archive.ics.uci.edu/ml/)

Typical structure of a DA report

Typically to be submitted to the project sponsor.

1) Abstract (a brief and accessible description of the project)
2) Introduction
3) Methods that were used for data acquisition and processing
4) Results that were obtained (put intermediate results in an appendix)
5) Conclusion
6) Appendices

Appendicies should include:

- ✓ All reproducible code used to process the data: well-commented scripts that can be executed without any command-line parameters and user interaction.
- ✓ The raw data: any data file that is required to execute the code in a reproducible way, unless the file has been provided by the data sponsor and has not been changed.
- ✓ A README file typically explains the provenance of the data and the format of every attached data file.

Suggestion on Course Work

- ✓ Start looking for team-mates to work on the project (max = 3)
- ✓ What problem are you interested in working on ?
- ✓ What data you will work on ? Where from do you collect the data ?

Literature survey : googling, reading research papers (may be)

Start early !! Sooner the better !!

What next ?

Week2: 16th January

Lecture : 9:00 – 11:00

- ✓ Introduction to Python
- ✓ Numerical Python library – Numpy

Lab Classes: 14:00 – 16:00

- ✓ Programming related to python and numpy.

Week3: 23rd January	– Matplotlib, Pandas introduction
Week5: 6th February	– Pandas in depth
Week6: 13th February	– Machine learning with Scikit-learn
Week7: 20th February	– Machine learning with Scikit-learn