Data Analytics

ECS684U / ECS784U / ECS784P

Part II: Machine Learning with Python

TIME-TABLE

Lectures/Tutorials:

- Friday: 9:00 – 11:00 AM
- Venue : Maths MLT

Lab classes:

- Friday 11:00 – 13:00 PM.
- Venue : ITL 1st floor (Week 6-11)
- First lab on 28th February.

Staff:

- Lectures : Bhusan Chettri and Anthony Constantinou.
- Labs : Janice (Chau Yi Li,), Anna Li, Aiqi Jiang, and Ammar Yasir Naich.

Module lecturers

Anthony Constantinou

Bhusan Chettri

Lab demonstrators

Aiqi Jiang

Ammar Yasir Naich

Anna Li

Janice

The module is grouped into 2 parts

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

- Comprises of statistical approaches for Machine learning such as Bayesian networks.
- Week 1, 2, 3, 4, 5

Part-II : Machine Learning with Python – Bhusan Chettri

- Comprises 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 6, 7, 8, 9, 10

Week 11: Revision week!

Topics covered in Part II (week 6-10)

- 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: Matplotlib library
- Machine learning using Scikit-Learn
- An introduction to Deep Learning with Keras
- Complexity analysis
- Case study of a Data Analytics project

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 components 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, diagrams and reports.

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, but cover just sufficient topics/materials 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 4 (max).
- More details in the coursework specification document.

Written exam (2 hours duration):

- Standard format, 4 questions (2 questions each from Part I and Part II) .

Data ?

Data ?

Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing.

Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs, images or other analysis tools.

Raw data (“unprocessed data”) is a collection of numbers of characters before it has been “cleaned” and corrected by researchers. Raw data needs to be corrected to removed outliers or obvious instrument or data entry errors (e.g. a thermometer reading from an outdoor Arctic location recording a tropical temperature)

Source: https://en.wikipedia.org/wiki/Data

Simple definition: *Anything that exists, represents facts, figures is a data. An image of a dog; numbers from 1 – 10 are examples of data.*

Data ?

A world of data – “Big Data”

* Data is produced and collected on a massive scale world-wide.

How much data do we create every day?

https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#5b33090660ba

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, Twitter, YouTube etc)

- Snapchat users share 527,760 photos
- Users watch 4,146,600 YouTube videos
- Instagram users post 46,740 photos

Check this link: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#5b33090660ba

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, Twitter, 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.

Types/classes of data

- Categorical: can be grouped or categorized. E.g. gender
- Nominal: categorical unordered data.
- Ordinal: categorical ordered data. E.g. educational experience.
- Interval: categorical ordered data where the size of the space between categories is the same. E.g. hike in salary with same increment !

- Numerical : integer or floating point numbers. E.g: 12, 13.54
- Discrete : which can take only set of values. E.g. No. of people in a classroom
- Continuous : values which can take any numeric value often in a specific range. E.g: person's height (within the range of human heights)

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.

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.)

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 ?

Stages /Phases in Data Analytics

1. Problem definition
2. Data source identification, selection and extraction
3. Data cleaning and transformation
4. Data exploration
5. Choosing modelling approach
6. Model development
7. Model validation/test
8. Visualization and interpretation of results
9. Deployment of the solution

Problem Definition

- Must have a clear understanding of the problem definition and the end goals.
- What do you want to address?
- Why do you want to address?
- Challenges.

Data source identification

- Identifying source of data for your problem.
- There can be multiple sources of data !

Data source selection

- We must ensure that the data source selected is “good” and reliable.
- How do we know if a data source is of “good quality?”

Data source identification

- Identifying source of data for your problem.
- There can be multiple sources of data !

Data source selection

- We must ensure that the data source selected is “good” and reliable.
- How do we know if a data source is of “good quality?”

Desirable qualities of a good data source.

- Less redundant.
- Does not have many missing values/incorrect/wrong values.
- Dataset is uniform and same through-out. Simplifies processing.
- Less imbalanced class data. E.g. For all n-classes there is equal amount of data for each class to avoid biases.
- Diverse – representative of a population.

Data source identification

- Identifying source of data for your problem.
- There can be multiple sources of data !

Data source selection

- We must ensure that the data source selected is “good” and reliable.
- How do we know if a data source is of “good quality?”

Data extraction

- Getting data from the data sources into your framework for model building !
- Python has good libraries for doing this from variety of sources (MySQL, NOSQL,excel etc.).

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.

Data exploration

- Summarizing data through data visualization
- Grouping data
- Exploration of relationships between the various attributes
- Identification of patterns and trends

Modelling approach and model development

What is a model?

What is a Model?

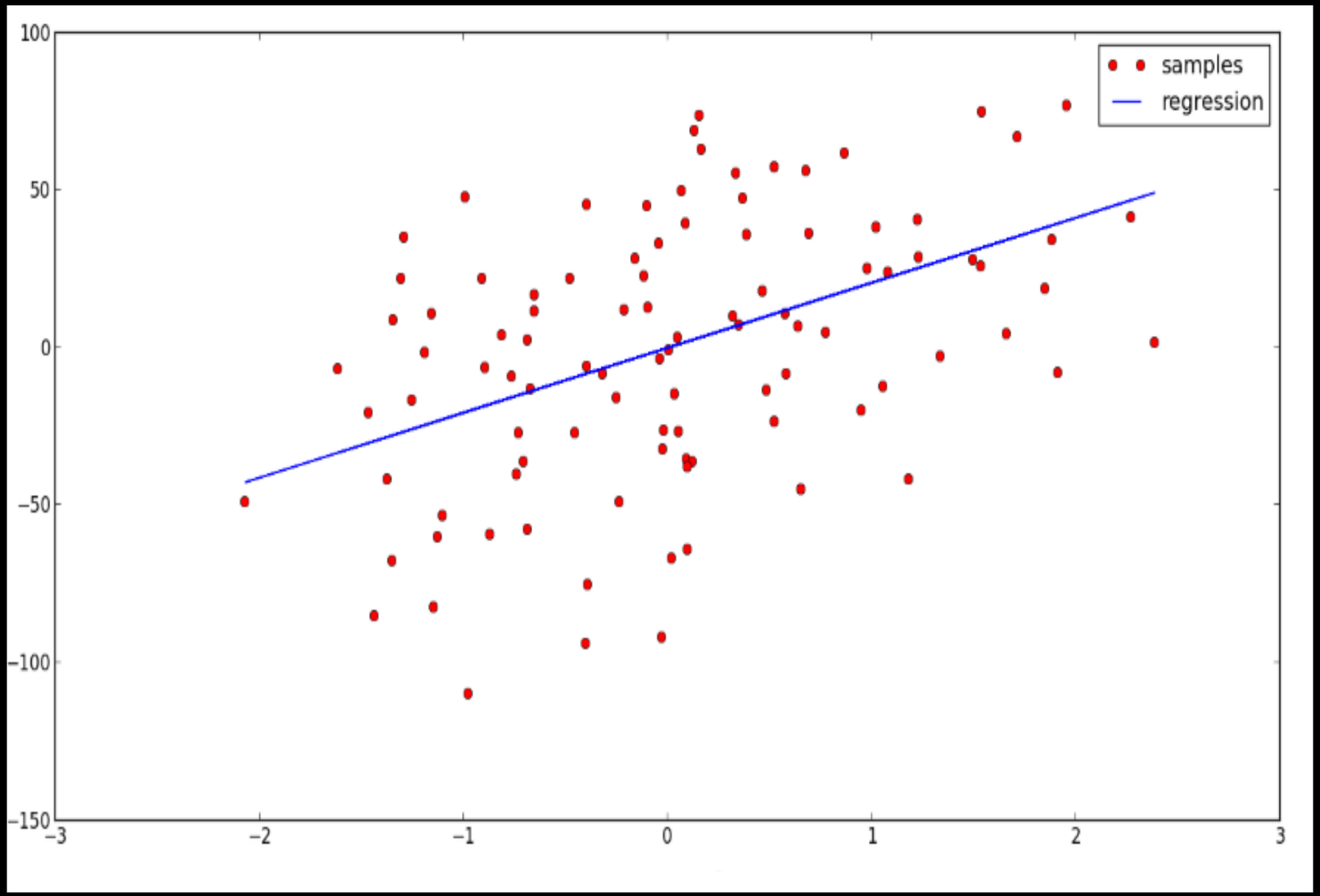
Data + Algorithm = Model

Example of a simple model:

$$Y = MX + C$$

- Linear regression model with one variable.
- Input to the model = X
- M and C are the parameters learned during training !
- Y is the output from the model.

Output (y Values)



Input features/values (x values)

An example of a linear regression model. Blue line represents the model.

Modelling approach and Model development

Modelling approach and model building can be classified based on the task at hand. Some common types include:

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

Model validation and testing

- ✓ The model we trained must show good performance when deployed in real time (unseen data).
- ✓ For this, we perform validation using a small portion of the data – called validation data in order to make sure that the model is generalising well and not just learning noise in the training data (a case of model overfitting !)

- ✓ 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 evaluation our model on the test set for final evaluation and deployment !

Interpretation of results

- ✓ Before deploying the model, we may want to perform some additional analysis on the results obtained on the validation and test datasets
- ✓ This may include analysing results in a tabular form, histogram analysis, different types of plots (pie chart, bar chart etc.) and other visualisations !

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

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

✓ 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/)