CM535 Data science development

Week 1 : Introduction

Benjamin Lacroix
b.m.e.lacroix@rgu.ac.uk

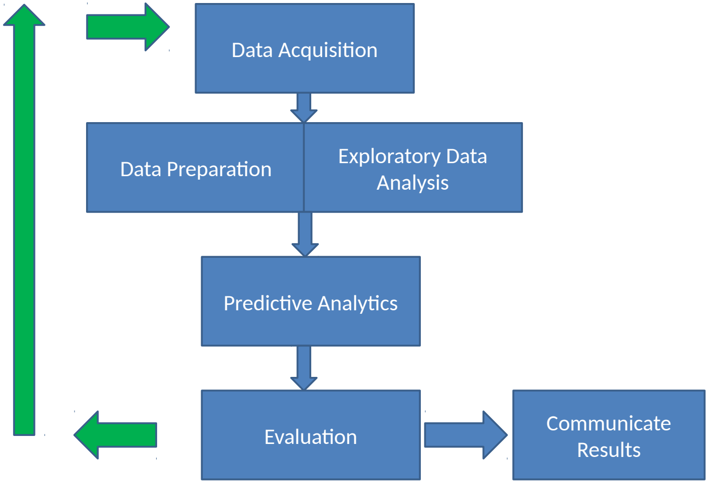
January 28, 2020

Learning Outcomes

- Discuss the main concepts and tools for a data science project.
- Load, explore, model and visualise data using off-the-shelf tools and packages.
- Report data science results to a wider audience by tailoring them at different levels of detail.
- Design, implement and evaluate a data science product that addresses a given data problem.

=> **Lead your own data science project**

Data pipeline



Data Acquisition

Data Preparation

Exploratory Data Analysis

Predictive Analytics

Evaluation

Communicate Results

Data acquisition

Where to find data to play around?

- Generate it yourself.
- Crawl the web:
  - UCI Machine Learning Repository `https://archive.ics.uci.edu/ml/index.php`
    - great for machine learning
    - clear presentation of the dataset
  - GitHub Awesome Public Datasets `https://github.com/awesomedata/awesome-public-datasets`
    - A bit of everything form different sources
  - Kaggle `https://www.kaggle.com/`
    - ML competitions
- "Scrape" specific web pages as permitted. An API is often provided

Target data format

Standard data format for machine learning algorithms and multivariate analysis

	Feature 1	Feature 2	Feature 3	Feature 4
Instance 1	3	1	Red	0.125
Instance 2	7	0	Blue	0.55
Instance 3	2	0	Red	0.99
Instance 4	15	1	Yellow	1.0
Instance 5	6	1	Green	0.3

Plus a **Code book** that:

- Describes the data
- Explains every feature

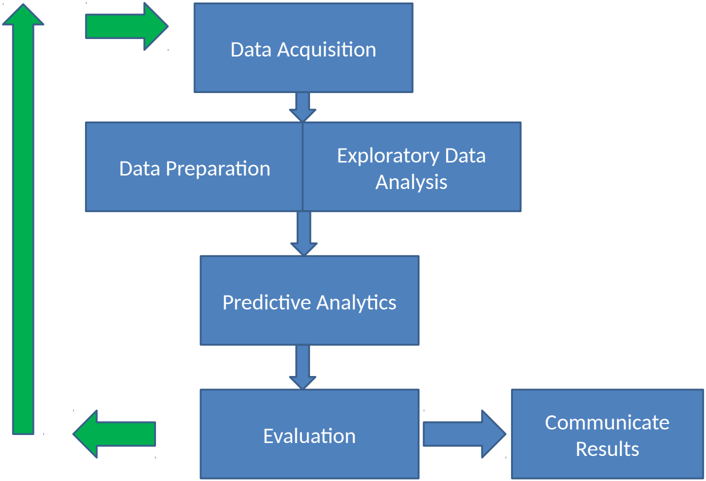
Target data format: Example

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.00	6.00	160.00	110.00	3.90	2.62	16.46	0.00	1.00	4.00	4.00
Mazda RX4 Wag	21.00	6.00	160.00	110.00	3.90	2.88	17.02	0.00	1.00	4.00	4.00
Datsun 710	22.80	4.00	108.00	93.00	3.85	2.32	18.61	1.00	1.00	4.00	1.00
Hornet 4 Drive	21.40	6.00	258.00	110.00	3.08	3.21	19.44	1.00	0.00	3.00	1.00
Hornet Sportabout	18.70	8.00	360.00	175.00	3.15	3.44	17.02	0.00	0.00	3.00	2.00
Valiant	18.10	6.00	225.00	105.00	2.76	3.46	20.22	1.00	0.00	3.00	1.00
.											
.											
.											

Features:

- mpg: Miles/(US) gallon
- cyl: Number of cylinders
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (1000 lbs)
- qsec: 1/4 mile time
- vs: V/S
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

Data pipeline



Data preparation and Exploratory Data Analysis

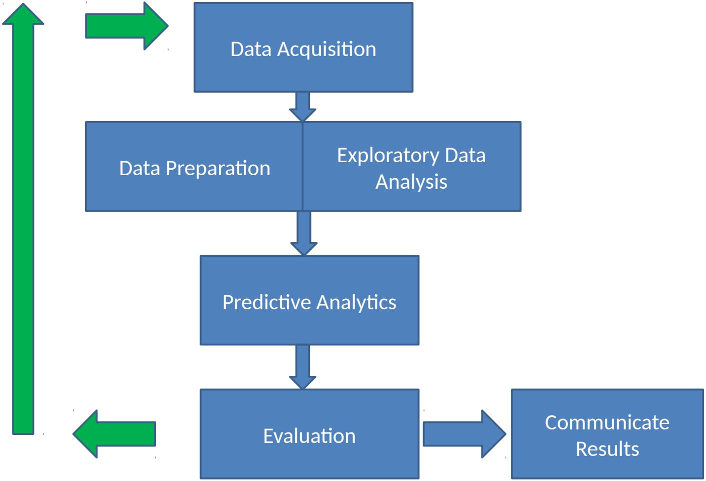
Data preparation:

- Data cleaning (missing/incorrect values).
- Data wrangling: putting the data in the right format for further analysis.
- Data transformation: constructing useful features from raw data.
- Feature selection.

Exploratory Data Analysis:

- Preliminary examination of data.
- Uses descriptive analytics
    - Descriptive statistics.
    - Data visualisation.
- Aims at understanding the data.
- Check for potential problems in the data

Data pipeline


Data Acquisition

Data Preparation

Exploratory Data Analysis

Predictive Analytics

Evaluation

Communicate Results

Predictive analtics

Supervised learning: Can past experience be used to predict the outcome of future instances?

Given a dataset X with m instances x_i defined by n features (called independant variables, predictors, explanatory variables) such that $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,n})$, associated to a response variable Y

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

Find function $h_\theta(X)$ that maps instances of X to an outout $y_i \in Y$

$$h_\theta(X) \approx Y$$

Where h in a model parameterised by Θ

Typical example

Does a patient have a heart disease?

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.30	0	0	1	1
37	1	2	130	250	0	1	187	0	3.50	0	0	2	1
41	0	1	130	204	0	0	172	0	1.40	2	0	2	1
56	1	1	120	236	0	1	178	0	0.80	2	0	2	1
57	0	0	120	354	0	1	163	1	0.60	2	0	2	1
57	1	0	140	192	0	1	148	0	0.40	1	0	1	1
56	0	1	140	294	0	0	153	0	1.30	1	0	2	1
44	1	1	120	263	0	1	173	0	0.00	2	0	3	1
52	1	2	172	199	1	1	162	0	0.50	2	0	3	1
57	1	2	150	168	0	1	174	0	1.60	2	0	2	1
67	1	2	152	212	0	0	150	0	0.80	1	0	3	0
44	1	0	120	169	0	1	144	1	2.80	0	0	1	0
63	1	0	140	187	0	0	144	1	4.00	2	2	3	0
63	0	0	124	197	0	1	136	1	0.00	1	0	2	0
59	1	0	164	176	1	0	90	0	1.00	1	2	1	0
57	0	0	140	241	0	1	123	1	0.20	1	0	3	0
45	1	3	110	264	0	1	132	0	1.20	1	0	3	0
68	1	0	144	193	1	1	141	0	3.40	1	2	3	0
57	1	0	130	131	0	1	115	1	1.20	1	1	3	0
57	0	1	130	236	0	0	174	0	0.00	1	1	2	0

Machine Learning

Different types of exercise:

- **Classifcation**: the output is a categorical
  - Binary classification
  - Multiclass classification ($\geq$3 classes)
- **Regression**: the output is quantitative/numerical $Y \in \mathbb{R}$

Machine Learning

Chose your model:

- Linear/Logistic regression
- Case base reasoning (KNN)
- Decision trees
- Ensemble methods (i.e. Random forrest)
- Kernel methods (i.e. SVM)
- Neural network
- Fuzzy set systems
- ...

Tune your model: Most models require certain parameters to be set a-priori and can have a strong influence on the performance of the model. Tuning usually requires testing different combination of those parameters.

Evaluation

- Evaluate the accuracy (or other performance metrics) of the prediction of a given model.
- Benchmark this accuracy against other models.
- Understand the performances obtained, what goes wrong? what goes well?
- What is the level of interpretability of the chosen model?
- Does the model provide some sort of insight on the performance?
    - how is the prediction made?
    - what are the most important features?

Communication

- Communicating results is key in a data science project.
- Data science is all about getting knowledge out of data.
- An analysis is worthless if unintelligible.
- Domain experts must understand your reasoning and conclusions.
- Need to know how to sell your product.

Teaching plan

Week	Date	Topic
1/21	28/01	Introduction to the module and R
2/22	04/02	Data preparation
3/23	11/02	Exploratory Data analysis
4/24	18/02	Linear regression
5/25	25/02	Logistic regression and Linear Discriminant Analysis
6/26	03/02	Variable subset Selection and PCA
7/27	10/03	SVM
8/28	17/03	Random forrest
9/29	24/03	Unsupervised learning
30		EASTER BREAK
10/31	07/04	Machine Learning Evaluation + Coursework
11/32	14/04	Coursework

The coursework

Put yourself in the skin of a data scientist who has been given a dataset. Your task will be to follow the data science process:

- Pick a dataset
- Data preparation
- Exploratory analysis
- Supervised learning experiment (classification or regression)
- Evaluation
- Presentation of results

Up next

Today's lab in N533

- Introduction to R
  - Data structures in R (vectors, matrices and data frames)
  - Manipulating data structures in R
- Introduction to Latex and Sweave

Next week:

- Data preparation/data cleaning.