Applied Data Science Coursework

Ian Nabney

The goal of this coursework is to give you experience of the whole lifecycle of developing a data science application. The learning outcomes for the coursework are those of the unit:

Acquire a working knowledge of practical data science, applied to real world problems. - Be able to start from raw data and deliver a representation allows a better understanding of the topics in the data.

* Have experience of using software tools for data pre-processing and management.
* Acquire first-hand experience in specific techniques for data storage.
* Understand the differences between different visualisation strategies to efficiently explore the data.
* Have learnt how to present and interpret data to/for a non-technical audience.
* Be able to share data under privacy constraints.
* Have practiced teamwork and time management.

Your tasks in this coursework are

* to follow a sound data mining application development process (for example, see the CRISP-DM standard) taking due account of privacy and other ethical issues;
* to develop models that will generalize well to new data;
* make a presentation on your work;
* to write a clear report on your findings.

The oral presentation will be delivered in the penultimate week of the teaching block (i.e. on the 29th April and the 1st May). Each group will present for 10 minutes with two minutes for questions. You will also get written feedback on the presentation which should help you to improve your report. The presentation is worth 10% of the unit, with the report counting for the other 90%.

Your report should contain the following sections (modified where appropriate for the needs of the particular problem you are studying:

* Abstract. A brief description of the key points in the report.
* Introduction. The background of the problem.
* Data Preparation. What data manipulation was necessary to create a dataset for analysis.
* Data Exploration. What you learned from your initial analysis of the data.
* Modelling. Which models you applied, their comparative performance, and a justification for your choice of the best model. You should include a brief description of the algorithms used (referencing the scientific literature where appropriate).
* Conclusion. What you have learned about the problem and data science from doing the coursework.

The assessment criteria are:

* Problem understanding: how well you have explained the goals of the project, taking account end-user requirements where appropriate. (10 marks)
* Data wrangling and exploration: care taken over extracting and manipulating the data; insights gained through the initial exploration of the data. (20 marks)
* Data modelling and evaluation: appropriateness of modelling approach; systematic use of statistical and machine learning methods; effectiveness and insight of the evaluation. (50 marks)
* Conclusions: what the problem owner should learn from your analysis. (10 marks)
* Presentation: fluency and coherence of the written text; quality of images and graphics used. (10 marks)

**Individual marks**

Each report will be marked using the criteria described above and this will be used to set an overall mark for each group. Individual marks will be determined based on this overall mark (so it is in everyone’s interests to collaborate well) with deviations based on each individual’s contribution. This will be assessed partly by your attendance at meetings with the group supervisor and partly on the basis of a 1-page reflective document that describes your role in the group, what your achievements were, and what you have learned from the course.

**Projects**

**Project 1**

**Geographic sentiment distribution of trending social media topics.**

In this task you will gather and analyse data from a social media network (Twitter suggested) to look at the geographic distribution of sentiment (i.e. emotional reaction) to a particular news item. This could be carried out at national (i.e. UK) or international level. Possible topics might include the EU referendum, major sporting events (such as the Olympic Games), gun control (in the USA), etc. Alternatively, instead (or in addition to if you are very ambitious), you could analyse the temporal changes in sentiment (i.e. how the sentiment changes over time).

**Project 2**

**Gender Equality**

Gender equality means that women and men and girls and boys enjoy the same rights, resources, opportunities and protections. Investments in gender equality contribute to lifelong positive outcomes for children and their communities and yield considerable inter-generational payoffs, as children’s rights and well-being often depend on the rights and well-being of women. UNICEF maintain a large number of datasets relating to a range of development topics, those related to gender can be found by exploring this site:

<https://data.unicef.org/topic/gender/overview/>

Your task is to build visualisations and predictive models to tell a compelling story about gender equality.

**Project 3**

**Understanding the impacts of loneliness.**

**End User: Jasmine Grimsley, Senior Data Scientist at the ONS Data Science Campus**

The Office for National Statistics (ONS) has developed a [loneliness index](https://datasciencecampus.ons.gov.uk/developing-a-loneliness-prescription-index/" \t "_blank) using open prescription data which is available at the MSOA (Middle layer Super Output Area in the ONS coding system) level across England. These data also provide information to identify MSOA’s that are within geographical clusters where the loneliness index is high or low. <https://github.com/matthewgthomas/loneliness/> Possible research questions and additional research data can be found here <https://www.bristol.ac.uk/golding/get-involved/competitions/loneliness-and-movement-for-education-competition/> That competition focused on the effects of loneliness on education, but other issues could be considered (e.g. economic outcomes or health/wellbeing).

**Project 4**

**Image classification on historical maps.**

**Data Owner: Yanos Zylberberg, Senior Lecturer in Economics**

The data consists of historical maps as well as some labelled pixel locations of chimneys. We are trying to digitize the maps and our approach to this is first to try to detect certain features (chimneys, trees), and hopefully next to recognize polygons and text. So, for the coursework, there are a number of different possible tasks. The most obvious would be to test different off the shelf image recognition on the labelled chimneys, but other possibilities include applying OCR to the street names, or detecting polygons (buildings, etc).

This link <https://uob-my.sharepoint.com/:u:/g/personal/in17746_bristol_ac_uk/EfyfSXsvHMhBgjFRNBxjwusBP0qJbv2wt10b6ncumQmsJg?e=Oj7VfA> is to a zip file containing

* raw data (Raw/): raw map tiles (geo-referenced, but this is not crucial) for Burnley and Oldham,
* locations of chimneys (Training/): locations of each chimney in each tile (Training/Locations, txt files); red marks on chimneys (Training/Marks, images).

All chimneys in the raw data are thus identified (about 500 in total).

**Project 5**

**Mental health and screen time.**

**Data Owner: Natalie Thurlby, Data Science Specialist, Jean Golding Institute**

Does the amount of screen time (using phones, computers, and tablets) a person spends at age 16 affect their levels of depression and anxiety at age 18? Data and data dictionary can be found on Blackboard on the projects tab.

**Project 6**

**Early Prediction of Sepsis from Clinical Data**

Sepsis is a life-threatening condition that occurs when the body's response to infection causes tissue damage, organ failure, or death. In the U.S., nearly 1.7 million people develop sepsis and 270,000 people die from sepsis each year; over one third of people who die in U.S. hospitals have sepsis (CDC). Internationally, an estimated 30 million people develop sepsis and 6 million people die from sepsis each year; an estimated 4.2 million newborns and children are affected (WHO). Early detection and antibiotic treatment of sepsis are critical for improving sepsis outcomes, where each hour of delayed treatment has been associated with roughly an 4-8% increase in mortality (Kumar et al., 2006; Seymour et al., 2017). To help address this problem, clinicians have proposed new definitions for sepsis (Singer et al., 2016), but the fundamental need to detect and treat sepsis early still remains, and basic questions about the limits of early detection remain unanswered.

This formed the 2019 Physionet Computing in Cardiology Challenge. The goal of this project is the early detection of sepsis using physiological data, which can be found, together with more background information, here <https://physionet.org/content/challenge-2019/1.0.0/>

**Project 7**

**Activity Recognition with Multimodal Sensor Data**

The task of this project is to predict aspects the activities of residents within a smart home based only on observed sensor data. Sensor data are obtained from the following three sensing modalities:

* A wrist-worn accelerometer
* Video + Depth (RGB-D)
* Passive environmental presence sensors

Data, metadata, and other information can all be found at this server.

<https://data.bris.ac.uk/data/dataset/8gccwpx47rav19vk8x4xapcog>

**Project 8**

**European Productivity**

This dataset is provided by the EU Open Data portal <https://ec.europa.eu/eurostat/tgm/graph.do?tab=graph&plugin=1&pcode=tesem160&language=en&toolbox=sort> The task is to analyse trends in this data, geographical and other relationships between outcomes. There are other datasets (such as that for GDP) on the website that may help you enrich your analysis.

**Project 9**

**Housing Statistics Analysis.**

the Government Statistical Service (GSS) has launched a new housing statistics interactive tool, which will help users explore the landscape and range of housing, planning, homelessness and rough sleeping statistics produced by Government. Your task is to use this data to illuminate a question in housing policy: affordable housing, homelessness, or some other social issue. Note that housing policy is a ‘devolved’ matter, which means that that each of the four countries in the UK (England, Wales, Scotland, and Northern Ireland) legislates and collects statistics based on its needs.

<https://gss.civilservice.gov.uk/dashboard/tools/housing-and-planning-statistics/filter-table.html>

**Project Allocation**

|  |  |
| --- | --- |
| Group | Project |
| 1 | 2 |
| 2 | 1 |
| 3 | 1 |
| 4 | 9 |
| 5 | 8 |
| 6 | 2 |
| 7 | 9 |
| 8 | 4 |
| 9 | 3 |
| 10 | 5 |
| 11 | 5 |
| 12 | 4 |
| 13 | 3 |
| 14 | 6 |
| 15 | 8 |
| 16 | 6 |
| 17 | 7 |