Data Analytics

ECS784P

Coursework specification 2017-18

Overview

The final project should represent significant original work applying data science techniques to an interesting problem. Projects should be done in groups of up to 4 people. There will be support from the module tutor/demonstrator during the coursework and you should make sure that you run at least one draft of the project plan by them before embarking on the detail of the project. This might be done face to face in the lab, or by email.

You should address a data-related problem in your professional field or a field you're interested in. Pick a subject that you're passionate about; if you're strongly interested in the subject matter it'll be more fun for you and you'll produce a better project!

Data sources

Using public data is the most common choice. If you have access to private data, that's also an option, though you'll have to be careful about what results you can release. Some sources of publicly available data are listed at the end of this document.

Project Deliverable

The deliverable from the project takes the form of a project report. Reports should be written with a technical audience in mind. It should be concise and clear, adopting the same style you would use in writing a report for management in industry or commerce, rather than a discursive essay style.

Here are the components you should aim to cover in your project report:

- Problem statement and hypothesis.
- Description of your data set and how it was obtained.
- Description of any pre-processing steps you took.
- What you learned from exploring the data, including visualizations.

- How you chose which features to use in your analysis.
- Details of your modeling process, including how you selected your models and validated them.
- Your challenges and successes.
- Possible extensions or business applications of your project.
- Conclusions and key findings.

Appendices

You should create appendices to your report that include:

- Code - commented Python scripts or any other code you used to develop your project
- Data - the data you used for your project, or, if this is too large, a pointer or pointers to your data sources.

Report Length/ Number of words

- The report shall have a minimum of 5000 and a maximum of 7000 words including the references. Please note that this does not include the content of your report appendix.
- Details of all project members need to be provided in the cover page.

Important Dates

- **Release date: 23rd January 2018**
- **Topic submission: 8th February 2018**
- **Deadline: Friday, 23rd March 2018**
- **Late Submission: 28th March 2018 (Penalty cost. After this date the online system will not accept submission)**

Example project plan and timetable

February 1st: Question and Data Set

What is the question you hope to answer? What data are you planning to use to answer that question? What do you know about the data so far? Why did you choose this topic?

Example: We're planning to predict passenger survival on the Titanic. We have Kaggle's Titanic dataset with 10 passenger characteristics. We know that many of the fields have missing values that some of the text fields are messy and will require cleaning, and that about 38% of the passengers in the training set survive. We chose this topic because we are interested in the history of the Titanic.

February 8th: Deadline for Topic Changes

You may discover during your data exploration that you don't have the data necessary to answer your project question. Therefore, you may decide to change the research question to address in the project. For this, you can consider 8th February as your deadline to decide and finalize the topic.

My advice is to spend initial time wisely doing some research on the data sources depending upon the problem you are trying to address. Researching appropriate data set to use in the project is also a part of your coursework. Various data sources links have also been provided for reference at the end of this material.

February 20th: Data Exploration and Analysis Plan

What data have you gathered, and how did you gather it? What steps have you taken to explore the data? Which areas of the data have you cleaned, and which areas still need cleaning? What insights have you gained from your exploration? Will you be able to answer your question with this data, or do you need to gather more data (or adjust your question)? How might you use modeling to answer your question?

Example: We have created visualizations and numeric summaries to explore how survivability differs by passenger characteristic, and it appears that gender and class have a large role in determining survivability. We estimated missing values for age using the titles provided in the Name column. We created features to represent "spouse on board" and "child on board" by further analysing names. We think that the fare and ticket columns might be useful for predicting survival, but we still need to clean those columns.

We analyzed the differences between the training and testing sets and found that the average fare was slightly higher in the testing set.

Since we're predicting a binary outcome, we plan to use a classification method such as logistic regression to make our predictions.

March 9th: have produced your first draft

At a minimum, this should include:

- Narrative of what you have done so far and what you are still planning to do, ideally in a format similar to the format of your final project paper
- Code, with lots of comments.

Ideally, you would also include:

- Visualizations you have done

Appendix: Open Data Sources

Political and Government Data

Data.gov

http://data.gov

This is the resource for most government-related data.

Socrata

http://www.socrata.com/resources/

Socrata is a good place to explore government-related data. Furthermore, it provides some visualization tools for exploring data.

US Census Bureau

http://www.census.gov/data.html

This site provides information about US citizens covering population data, geographic data, and education.

UN3ta

https://data.un.org/

UN data is an Internet-based data service which brings UN statistical databases.

European Union Open Data Portal

http://open-data.europa.eu/en/data/

This site provides a lot of data from European Union institutions.

Data.gov.uk

http://data.gov.uk/

This site of the UK Government includes the British National Bibliography: metadata on all UK books and publications since 1950.

The CIA World Factbook

https://www.cia.gov/library/publications/the-world-factbook/

This site of the Central Intelligence Agency provides a lot of information on history, population, economy, government, infrastructure, and military of 267 countries.

Health Data

Healthdata.gov

https://www.healthdata.gov/

This site provides medical data about epidemiology and population statistics.

NHS Health and Social Care Information Centre

http://www.hscic.gov.uk/home

Health datasets from the UK National Health Service.

Social Data

Facebook Graph

https://developers.facebook.com/docs/graph-api

Facebook provides this API which allows you to query the huge amount of information that users are sharing with the world.

Topsy

http://topsy.com/

Topsy provides a searchable database of public tweets going back to 2006 as well as several tools to analyze the conversations.

Google Trends

http://www.google.com/trends/explore

Statistics on search volume (as a proportion of total search) for any given term, since 2004.

Likebutton

http://likebutton.com/

Mines Facebook's public data--globally and from your own network--to give an overview of what people "Like" at the moment.

Miscellaneous and Public Data Sets

Amazon Web Services public datasets

http://aws.amazon.com/datasets

The public data sets on Amazon Web Services provide a centralized repository of public data sets. An interesting dataset is the 1000 Genome Project, an attempt to build the most comprehensive database of human genetic information. Also a NASA database of satellite imagery of Earth is available.

DBPedia

http://wiki.dbpedia.org

Wikipedia contains millions of pieces of data, structured and unstructured, on every subject. DBPedia is an ambitious project to catalogue and create a public, freely distributable database allowing anyone to analyze this data.

Freebase

http://www.freebase.com/

This community database provides information about several topics, with over 45 million entries.

Gapminder

http://www.gapminder.org/data/

This site provides data coming from the World Health Organization and World Bank covering economic, medical, and social statistics from around the world.

Financial Data

Google Finance

https://www.google.com/finance

Forty years' worth of stock market data, updated in real time.

Climatic Data

National Climatic Data Center

http://www.ncdc.noaa.gov/data-access/quick-links#loc-clim

Huge collection of environmental, meteorological, and climate data sets from the US National Climatic Data Center. The world's largest archive of weather data.

WeatherBase

http://www.weatherbase.com/

This site provides climate averages, forecasts, and current conditions for over 40,000 cities worldwide.

Wunderground

http://www.wunderground.com/

This site provides climatic data from satellites and weather stations, allowing you to get all information about the temperature, wind, and other climatic measurements.

Sports Data

Football dataset

http://www.football-data.co.uk/

Pro-Football-Reference

http://www.pro-football-reference.com/

This site provides data about football and several other sports.

Publications, Newspapers, and Books

New York Times

http://developer.nytimes.com/docs

Searchable, indexed archive of news articles going back to 1851.

Google Books Ngrams

http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

This source searches and analyses the full text of any of the millions of books digitized as part of the Google Books project.

Musical Data

Million Song Data Set

http://aws.amazon.com/datasets/6468931156960467

Metadata on over a million songs and pieces of music. Part of Amazon Web Services.