`ENVS615` - Analysis of Human Dynamics

- Dani Arribas-Bel [`@darribas`]

Schedule

Day 1

a. **10:00-11:00** | Introduction: *data, data, data* (`block_1a`)
b. **11:00-12:00** | The computational building blocks of data science (GDS Ch.1)

[Lunch Break]

c. **13:00-14:30** | Interacting with tabular data lab (Pt. I, `01_tabular_data`)

[Break]

d. **14:45-16:00** | Interacting with tabular data lab (Pt. II, `02_tabular_data_advanced`)

Day 2

a. **10:00-12:00** | Visualising tabular data lab (`03_tabular_data_viz`)

[Lunch Break]

b. **13:00-14:00** | Unsupervised learning (`block_2b`)

[Break]

c. **14:15-16:00** | Unsupervised learning lab (`04_unsupervised_learning`)

Day 3

a. **10:00-11:00** | Supervised learning
b. **11:00-12:00** | Supervised learning lab (Pt. I, `05_supervised_learning`)

[Lunch Break]

c. **13:00-14:30** | Supervised learning lab (Pt. II, `06_inference`)

[Break]

d. **14:45-15:30** | Supervised learning lab (Pt. III, `07_overfitting_cv`)
e. **15:30-16:00** | Assignment details + questions + pick your favorite

Day 4

1. **10:00-12:00** | Pick your favorite: Geo Vs APIs

[Lunch Break]

2. **13:00-14:30** | Data Science Studio

[Break]

3. **14:45-16:00** | Data Science Studio

Bonus

Data preparation available in `zzz_data_prep`

Collaboration

The module has a related Microsoft Teams team set up on the Liverpool cloud. You should have received an invite directly to your **Liverpool email** account but, if you have not, you can join the Team on the following link:

> https://teams.microsoft.com/l/team/19%3a1822ee5f66654039938ba7b7f7ef8715%40thread.skype/conversations?groupId=910b24fd-6449-42b5-b6f3-5f416c0b9cd2&tenantId=53255131-b129-4010-86e1-474bfd7e8076

IMPORTANT: you will need to be logged in on your Liverpool account (not your own university if you are not a Liverpool student).

Assessment

Key information:

- Type: Coursework
- [Equivalent to 5,000 words] Up to five figures and three tables + code + comments + up to 2,000
- Chance to be reassessed
- Due on **March 9th at 2pm**
- Electronic submission only. Static HTML with NO interactive cells

This module is assessed through a *computational essay*. To complete it successfully, you will need to demonstrate aptitude in at least three areas:

1. Data audacity
2. Python data skills
3. Machine learning and inference literacy

These translate in the following components of the computational essay:

1 - Find, prepare & explore a dataset

Find a dataset you are excited about and that meets the following characteristics:

- It contains several characteristics (features) for a number of observations (samples)
- At least two characteristics are continuous and at least two are categorical
- You can think of ways in which clustering the observations based on their characteristics could tell an interesting story
- You can imagine a situation in which one of the continuous characteristics can be explained in a supervised model as a function of some of the other characteristics

NOTE: please discuss with Dani the choice of dataset before the course finishes

With the dataset at hand:

1. Prepare it for analysis
2. Explore the dataset visually, identifying interesting patterns

2 - Unsupervised learning

Perform a clustering exercise & analyse the results. You are expected to try several clustering models, choose a preferred one, and present a critical argument about why that is your choice. To build your argument, you may rely on graphics, performance scores, and substantive reasoning. Demonstrate that you understand not only how the mechanics of the algorithms work but that you are able to translate those into an applied context to make sense of data.

3 - Supervised learning

Finally, build a predictive model based on linear regression and:

- Interpret the coefficient
- Evaluate its predictive performance both with and without cross-validation
- Reflect on the differences between assessing the performance of a model cross-validating and not.

Similarly to the previous point, demonstrate that you both understand the workings of the algorithms and techniques but also how you can make the most of it to learn about your data. Critical thinking is critical.

Further materials

A living list of further materials where you can continue learning is updated at:

http://darribas.org/gds19/further_resources.html

Infrastructure

This course requires the following libraries installed:

- jupyterlab
- pandas
- scikit-learn
- seaborn
- statsmodels

You can install these through most modern Python package managers (e.g. conda). If you have administrative rights on your machine and are running either Windows 10 Pro, macOS, or Linux, a containerised platform is advised. The following two are good options:

- jupyter-scipy-notebook: the official Jupyter container for data science in Python.
- gds: a more comprehensive, geo-focused stack for (geographic) data science in Python and R.

You have instructions to install and run the containers at:

http://darribas.org/gds19/software.html