Data Analysis Workshop

Objective

Deliver a standalone workshop for approximately 2 hours which meets the needs of thesis students.

Proposal

The workshop could be broken into smaller themes which would roughly align with the sequence of tasks students face when writing their thesis. Principally, the sub themes could be:

- Best practice for collecting data;
- Best practice for storing and manipulating data;
- Tips, tools, and advice for analysing data;
- How to correctly perform hypothesis testing, and the pitfalls of inference;
- Basic machine learning techniques (principally regression)
- Data compression techniques (pca);
- Questions and answers (where to find further information)

Each of the themes would be comprised of theory (where applicable), demonstrated technique, and working examples in which students would be given the opportunity to work with data in a test environment.

Sections

Best Practice for Collecting Data

- Successful creation of a control to contrast experimental data
- Avoiding bias when collecting data (*though this may be less important for hard science*)
- Having an understanding of the domain of the model that you want to test to avoid extrapolation

Best Practice for Storing and Manipulating Data

- Potential to demonstrate to students the best way in which data can be stored - this should be done with an observation in each row, and variables restricted to columns. Redundancy should be encouraged to provide maximum
- Data manipulation should be done in a spreadsheet application. This provides the student with a handy way to visualise the data, and address irregularities.
- *There may be scope here to talk about addressing model outliers.*
- Students given a best practice for storing data. Typically from spreadsheets they could use .csv files - which will provide the highest level of accessibility when loading their data into an analysis software (e.g. MATLAB, python, R)

Tips, Tools and Advice for Analysing Data

- *What types of analysis are students trying to run?*
- The following are a list of ideas that students typically get wrong when performing statistical analysis:
    - Incorrectly formatting hypotheses
    - Choosing the Incorrect statistical tool to test hypotheses
    - Incorrect formats for reporting of the statistical tests performed
    - Extrapolating models and the inaccuracies that come with this
- A review of commonly used statistical tools.
- Posing questions, and how to convert them to a format suitable for data analysis, and which tool to use for the job (this would be guidance only).
- Demonstrating to students what constitutes a good graph, using pictures to tell a story, and how to select the right picture for the story that they want to tell.

How to correctly perform hypothesis testing

- An entire section could be devoted to showing students how to perform hypothesis testing
- Start with talking about the parameters that students typically form hypotheses around (e.g. mean and standard deviation). Provide many examples of this - easiest to learn by way of example. (Students often find these concept abstract)
- Talk about the statistics that are important when performing statistical tests
- One tail, two tail testing - selecting the appropriate hypothesis for a given situation
- Using standard normal model and t-model for testing hypotheses on the mean of a distribution.
- Variance testing using the F distribution to compare two variances
- Providing students with the correct language for reporting the results of their hypothesis testing.

Basic machine learning techniques (principally regression)

- Talk about the different software packages that are available to students for this type of analysis (open source is always good).
- Demonstrate to students how to load a data set into a software (using csv)
- Demonstrate to students how to perform a simple multivariate regression.
- Cover the theory of what you can, and cannot do with simple multivariate regression models. (Good for providing statistical backing for ideas, bad for providing a predictive model).
- Explain the benefits of these models - adding parameters to the model to control for factors.
- Provide students with the correct language for describing the results of the model.
- When it is appropriate to use MVR and when it is not appropriate (e.g. heteroskedastic data and data which can present with autocorrelation/time series data)

Questions to help focus and drive this meeting

- What are the most common errors that students are making with their thesis papers?
- What types of data analysis are most commonly used amongst students?
- What are the desired learning outcomes for this session?
- How long does each section need to be? How many sections would be best to include?
- What level of sophistication do students have with their approaches to data analysis?
- What materials should be produced to support the session? Lecture slides, oral delivery, punctuated with hands on exerciese
- What should the structure of the session be? Theory hands on theory hands on … etc