INSTG083/INSTM083: Group Project Description
Regression Analysis

Set by: Luke Dickens

Important Notice

This assessment forms part of your degree assessment. All work must be done by group members:

- You must not collaborate or work with students outside of your group.
- You must not send or show other groups your work.
- You must not ask students outside of your group for help, or ask to see their work. As well as being against regulations, this is unfair to the other student concerned, since it may lead to them being accused of plagiarism.
- You must not seek help from friends, relatives or online discussion groups, other than the moodle forum for INSTG083/INSTM083.
- If you think any of the description of the task below is ambiguous or unclear, please post to the moodle forum, explaining what your concerns are, or raise it in person with your lecturer, Luke Dickens, or an INSTG083/INSTM083 lab demonstrator.
- If you are unsure of any of the above points, please post your concern to the moodle forum.

Finally, if there is any reason you think you cannot contribute fully to the group work in the alloted time, you should discuss your reasons with the INSTG083/INSTM083 lecturer, Luke Dickens at `l.dickens@ucl.ac.uk`.

1 Task Description

On the course website*, there is a data-file containing data relating to a selection of Portugese wines. Each row of the data corresponds to a different sample of wine, and includes information about the physicochemical properties of the wine, as well as a quality assessment based on sensory experience (the columns of the data). Your job, as a group, is to try to perform regression analysis on this data-set. The regression objective is to predict the quality score, based on the other attributes. You should implement, fit and evaluate a selection of regression methods you have encountered during the course to determine which perform best at this regression task.

A typical investigation like this would include:

- Some exploratory data analysis, to understand the data and how you will encode it for your models.

*The data file and a description file can be accessed here: `https://moodle.ucl.ac.uk/course/view.php?id=42289&section=2`

- The selection of a variety of methods to evaluate. This may involve: choosing which method types to evaluate, e.g. simple linear regression, kNN regression, linear models etc; which types of basis functions to explore; and how you are going to train these models.
- Implemention of all the necessary code to train and evaluate the selected models.
- Performing training and parameter selection within each model, as well as a final comparison across your main models. If you are splitting data, you should decide what proportions to use for training, validation and testing.
- An interpretation of your findings. This should reflect on the assumptions you have made for each model, its relative performance, and the certainty of your findings.

On completion, each group should submit one report describing this process, as well as a code archive (zip). When extracted and run, the code should recreate all computational processes, figures and results described in the report. Individual students should also submit a short report reflecting on what they learnt during the assignment, including a breakdown of work among group members.

2 Admin & Submission

This section contains information on what should be submitted for your group project assigment, and by whom.

2.1 Groups and Group Representatives

Each student has been assigned to a group and should work together in your group to produce a group report and supporting code. Each group should choose a group representative, who will be responsible for uploading the final submission of these two components. Individual group members must also submit an individual report to be submitted separately.

2.2 Some Rules and Advice

Reports will be assessed on their clarity, the appropriateness of methodologies & figures, and the relevance of the findings. Any models and methods should be correctly and clearly described and the evaluation procedures (e.g. cross-validation) explained. Students do not need to re-explain in depth material covered in lectures, but should give the intuition behind and constraints of each approach, as well as describing any specific choices such that an independent researcher could reconstruct their experiments, e.g. describing: the number of runs; how data was divided for training, validation & testing; how parameters were evaluated and selected and so on.

Figures should be clearly drawn, with text having an appropriate size, and content easily interpretable. Plot axes should be clearly labelled, and, where appropriate, plotting lines or marks given a legend. Appropriate choices should be made for whether to scatter, histogram, line plot, bar plot, or otherwise display data. If they are appropriate, error bars or confidence ranges can (and should) be included. All figures and tables should have clear captions, and salient features discussed within the body text.

Overall you should compare a small number of general models/complete methods. Groups of three students should have 3 or 4 general models; groups of four students should have 4 or 5 general models[†]. General models

[†] There is no obligation to analyse the maximum number of models allowed by your group size. It is quality, not quantity that counts.

can describe a complete method, e.g. least-squares linear regression; or a collection of choices which together describe a complete method, e.g. Bayesian regression on a linear model with rbf basis functions. Intermediate analyses should evaluate the best tuning parameters for each complete method, e.g. number or width of basis functions or regularisation parameter. Your final analysis should compare these complete methods, reporting the performance of the best parameter selection for that approach.

The group report should conclude with a discussion of which methods performed best on the data, the possible reasons for this, what assumptions were made, and how certain the students are of their findings.

Code will be assessed on whether it runs without errors, and should recreate all results/plots included in the report. You should not edit the data-file. Submitted code will be tested on data in the same format as provided on the website.

For the individual reports, students should reflect on what they have learnt during the group exercise. For this, you should focus on general concepts you have understood more clearly, and skills you have developed during this practical exercise. You should also describe how the work/effort for the project was distributed among group members.

2.3 When you are ready

When you are ready your group representative should submit the project report (as a doc, docx or pdf) and the code (as a zip archive) at the corresponding link provided in the assignment brief. Individuals should submit individual reports at the corresponding link, again provided in the assignment brief.