Master of Science (Business Analytics)

MIS41120: Statistical Learning

Practical Data Analysis Assignment

Due date and time: 9:00am, Monday 8th April, 2019.

Assessment weight: 25%.

1. Introduction

This team assignment will involve the team choosing a real or realistic dataset (preferably a freely available dataset), applying to it methods from this course, and analysing the results according to various criteria, *e.g.*, performance, accuracy, interpretability, efficiency, etc.

There are two main purposes in doing this assignment:

(*a*) to develop further your ability to investigate a dataset using an advanced tool such as R, so building on your practical work; and
(*b*) to develop your analysis and reporting skills in conveying the main results of your analysis as a written report.

The assignment will be done in teams of three. It will be graded according to the following criteria:

(*a*) Quality of your R code (including quality of comments);
(*b*) Quality of written report.

If the assignment is late, it will be penalised: it

- will lose 10% of the assignment mark if up to one week late;
- will lose 20% if between 1 and 2 weeks late; and
- *cannot* be accepted if more than 2 weeks late.

These are general UCD regulations, outside my control.

Also, please see the University policy on plagiarism.

Both of these are at `http://www.ucd.ie/registry/academicsecretariat/pol.htm`: look under L for "Late Submission of Coursework" and P for "Plagiarism and Academic Integrity", respectively.

2. Task

Return to Question 15 on page 126 at the end of Chapter 3 of the textbook ISLR (James et al, 2014). This asks you to use R for linear regression with least squares on the Boston dataset from the MASS library. Complete part (b) of this question:

> Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

You may need to review the earlier chapters of ISLR.

Do this using each of the following regularisation approaches:

(*i*) no regularisation
(*ii*) ridge regression (ℓ_2)
(*iii*) lasso (ℓ_1)
(*iv*) elastic net (ℓ_1 and ℓ_2).

Now carry out all of the above on the Boston dataset using two other learners:

- support vector machine (SVM) and
- multilayer perceptron (MLP).

That is, for each of SVM and MLP, use all four of: (*i*) no regularisation; (*ii*) ℓ_2 regularisation; (*iii*) ℓ_1 regularisation; (*iv*) ℓ_1 and ℓ_2 regularisation.

If a part of the ISLR Question 15(b) is not relevant to an extended methods, you can ignore that part.

Now do all of this (regression, SVM and MLP with all four regularisation approaches (*i*)–(*iv*)) on a publically available dataset with one output variable and at least 20 predictors (input variables). Explain your choice of dataset.

Where appropriate, use k-fold cross-validation (splitting into training and test sets k times) to estimate the model quality.

You may need to write some R functions for certain regularisation approaches. Make sure you comment these well. You might like to look at the R packages `sparseSVM` (`https://cran.r-project.org/web/packages/sparseSVM/sparseSVM.pdf`) or `penalizedSVM` (`https://cran.r-project.org/web/packages/penalizedSVM/penalizedSVM.pdf`) at CRAN. For neural networks you could try the R packages `snnR` and `brnn` or others.

In your report, comment on which methods were superior and — where possible — explain why. Did it depend on the dataset?

3. Deliverable

Submit two deliverables as described below.

The first deliverable is a written report on your work, in Word, Openoffice or pdf format. It has the layout:

(*a*) A standard cover page stating that this is all your own work, signed by all team members;

(*b*) A title page, containing
- title and handup date of assignment
- full name and student number

(*c*) At most ten pages of text containing your analysis and conclusion, no smaller than 10 point font.
- Include a URL link to the publically-available dataset you used. If this dataset is not too big, you can also include it in the zipfile (second deliverable — see below).

(*d*) Diagrams (which can be put at the end of the document) do not count towards the twn page limit.

The cover page and title page (these may be combined into one page) do not count towards the page limit.

The second deliverable is a zipfile of all the R scripts you used in the assignment. Each script must be self-contained; that is, it should run "out of the box" when I run it inside R or RStudio.

Both deliverables must clearly indicate all team members' surnames; name them according to the convention

`stl_Surname1_Surname2_Surname3_report.pdf` (or `.docx`, `.odt`, etc.) — written report and

`stl_Surname1_Surname2_Surname3_code.zip` — zipfile of R code (and possibly data)

Submit your deliverables through Brightspace. One team member should submit the report and another should submit the zipfile.

Also, as a backup, email both deliverables to `sean.mcgarraghy@ucd.ie`. CC all team members. The subject of the email must be your project name `stl_Surname1_Surname2_Surname3`