Master of Science (Business Analytics)

MIS41120: Statistical Learning

Practical Data Analysis Assignment

Due date/time: 5:00pm, Tuesday 28th July, 2020.

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:

(*i*) to develop further your ability to investigate a dataset using an advanced tool such as R, so building on your practical work in the class labs; and
(*ii*) 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, with one or two teams of two doing a reduced-size assignment if the class size is not evenly divisible by three. Please form groups yourselves. When you have formed your group, email the names and student numbers of all members to me at `sean.mcgarraghy@ucd.ie`. Make sure that you CC all team members and that the subject of the email is your project name `stl_Surname1_Surname2_Surname3`.

It will be graded according to the following criteria:

(i) quality of your R or Python code (including quality of comments);
(ii) quality of written report.

Also, please see the University policies on plagiarism, etc.

These are at `http://www.ucd.ie/registry/academicsecretariat/pol.htm`

2. Task

The computation for the following is most easily done in R, since that is what we have covered in tutorials; but, if you prefer, you may use Python.

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 $H_0 : \beta_j = 0$?

Now repeat this using each of the following regularisation approaches:

- ridge regression (ℓ_2)
- lasso (ℓ_1)
- elastic net (ℓ_1 and ℓ_2).

Now carry out all of the above using two other learners: support vector machine and multilayer perceptron, the standard artificial neural network (ANN). For each, use all three of ℓ_2, ℓ_1 and elastic net regularisation. If a part of the ISLR Question 15(b) is not relevant to an extended method, you can ignore that part.

Now do all of this (regression, SVM and MLP with all three regularisation approaches) on a publically available dataset with one output variable and at least 20 predictors (input variables).[1] Explain your choice of dataset.

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

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

[1] See §4 for some possible sources of data.

2.1. **Note regarding implementation.** The usual R `e1071` package does not easily support regularisation other than basic ℓ_2. The R package `sparseSVM` supports ℓ_1 and elastic net regularisation for linear kernels, but only for binary classification. The situation is similar for the `penalizedSVM` package. Thus you may need to convert a regression problem to a classification one to test these approaches: *e.g.*, if the numerical response is above a certain threshold T, assign class label 1; otherwise, assign class label 0. This is acceptable as a solution. Other packages you may like to look at include `kernlab` and `ksvm`, which do support vector regression, but only have ℓ_2 regularisation.

For ANNs there are a few R packages to look at such as `snnR` and `neuralnetwork` which supports both ℓ_1 and ℓ_2 regularisation.

You may need to write some R or Python code for certain regularisation approaches. Make sure you comment these well.

3. Deliverable

Submit two deliverables as described below.

The first deliverable is a written report on your work. This deliverable may be a Word, Openoffice or pdf file. It has the form:

(i) A standard cover/title page, containing
- title and handup date of assignment
- full name and student number of all team members
- a statement that this is all your own work, signed by all team members.

(ii) 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).

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

The cover/title page does not count towards the page limit.

The second deliverable is a zipfile of all the R or Python 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. Only one team member should do this, to avoid false positives on the plagiarism detector.

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

4. Sources of data

As a general principle, for an assignment like this, you do not want to spend a lot of time cleaning data. Thus, you should stick to datasets that are already cleaned. Also, for computational cost reasons, your chosen dataset should not have too many predictors (input variables) or observations

(instances). However, the number of observations should ideally be at least 5 times the number of predictors.

There are many free publically-available sources of data that you can use. Below are some suggestions, but there are many others.

- `http://mlr.cs.umass.edu/ml/` or `https://archive.ics.uci.edu/ml/index.php` the UCI ML repository, one of the oldest and best-known sites: most but not all datasets are clean
- `https://www.cso.ie/en/index.html` Irish CSO (Central Statistics Office)
- `https://www.ons.gov.uk` UK equivalent of CSO
- `https://www.ukdataservice.ac.uk` UK collection of social, economic and demographic data
- `https://www.data.gov` US Government open data
- `https://www.kaggle.com` User-uploaded data sets, also hosts model competitions
- `https://data.world` a repository of public datasets
- `https://fivethirtyeight.com` mostly political and economic data, the datasets are hosted at `https://github.com/fivethirtyeight/data`. A good site to look at for interpretation of data
- `https://opendata.socrata.com` big but less easy to find what you want, and the datasets may not be clean
- `https://www.quandl.com` mostly financial data, and some is not free
- `https://github.com/BuzzFeedNews/everything` some good datasets, others less so
- `https://www.reddit.com/r/datasets` the subreddit `/r/datasets` has a lot of interesting datasets, of variable levels of cleanliness
- `http://academictorrents.com` mostly datasets from academic papers
- `https://trends.google.com` tons of stuff here
- `http://networkrepository.com/index.php` Network data, perhaps less relevant for this project
- `https://sites.google.com/site/ucinetsoftware/datasets` Social Network data, again, maybe less relevant for this project

See also sites such as the following, which either point you to other sites, or can allow you to find sources of data:

- `https://toolbox.google.com/datasetsearch`
- `https://guides.library.cmu.edu/machine-learning/datasets` Carnegie-Mellon University, Pittsburgh, curated by Huajin Wang
- `https://medium.com/towards-artificial-intelligence/the-50-best-public-datasets-for-machine-learning-d80e9f030279`