Statistical Learning 2018 Coursework 1 – Prediction and Classification

January 21, 2019

Overview

The data (*reconv.csv*) in this task are derived from reoffending data collected across England and Wales.

- The variables in the data set are shown in Appendix A.

Each row in the data file represents one offender who was convicted in 1990 – the date of this conviction is termed the 'target date' and the main offence at that conviction is the 'target offence'. The offender's criminal history up to that point has been summarized, and a binary variable has been used to indicate whether the offender was reconvicted (for any offence) in the follow-up period for this data set (up to the end of 1999). Your task is to perform analysis of this data with the end goal of predicting re-offending, and attempting to understand which factors may impact this.

- **Coursework is due in on Monday of Week 16 (18th February 2019)**
- **This is an individual exercise!!**

Description of Task

The Sections below describe the tasks you will be expected to carry out during your analysis.

1. Data Manipulation and Exploratory Data Analysis

Split the dataset into a training set and a validation set. Pick an appropriate proportion of the data to keep in the validation (test) set. Check for class imbalance in training and test set. Perform some basic explotatory data analysis on the variables within the data-set. Consider (briefly) prior work to assess which variables domain experts consider important for the task. Some background to the task can be found here:

- https://modules.lancaster.ac.uk/mod/page/view.php?id=991449

Remember, to enable fair comparison the test and training datasets for each of the tasks below should be the same!

2. Tree Models

Create a full classification tree on the training set to determine which of the criminological variables are predictors of whether the offender is reconvicted or not by the end of 1999 ('reconv'). Choose an appropriate criteria and consider a pruning strategy in order to obtain a tree which is not too unwieldy but retains a reasonable accuracy. You may do this either via caret, or manually, however, you must describe the procedure used here. Describe the final tree and explain what variables affect reconviction, and how they do. Use the validation set to assess performance.

3. Logistic Regression

Fit a logistic regression to the training set data and examine the output in a similar way, assessing performance on the validation set. Describe the final model and explain what variables affect reconviction, and how they do. Again, use the validation set to assess performance.

4. Neural Networks

Fit a neural network to the training data, changing the number of hiden units. For simplicity, in this case only add one hidden layer to the network, however, you should vary the number of hidden nodes in the model. Pick your preferred model based on the cross-validated performance (not the validation set). Report the performance of your chosen model on the validation set.

5. Comparison and Discussion

Discuss your final prediction models for each of the above sections (2,3,4) comparing them to each other. You can use either the misclassification rate, accuracy or a likelihood based measure to do this (eg Deviance, AIC, or entropy). What variables contribute to each? Which appears to fit the best on the validation set? If this method were to be used in the courts to determine outcome at sentence, which would you recommend to the Ministry of Justice, and why? Which method do you recommend for giving highest predictive performance.

6. Predictions

Finally, I have held out a further sample of data from this data-set. You can find the covariates X for this data in the file, *"reconv_predict_no_label.csv"*, I have left out the reconviction outcome, you need to predict this. Create a set of predictions for this data for the outcome of 'RECONV'. Create your predictions in R using the model of your choice*, i.e as recommended in Task 5.* Your predictions should be created as a dataframe in R, with ***columns labeled 'OINUM' and 'RECONV'*** where id is the OINUM for the relevant individual and RECONV is your prediction. The prediction should be binary coded as "Yes", or "No" (not as 1/0 or anything else). Save your predictions as a .csv file using the command:

```
write.csv(your_prediction_df, file="student_id_predictions.csv")
```

Coursework Outputs

Report

You should write the above tasks up as a report

- Main report $\leq$ 4 pages of A4,
- Minimum font-size 10
- Single line spacing
- Margins not smaller than 2.5cm
- One extra A4 page of Appendices (i.e. extra figures/tables) is allowed

Feel free to structure the report (i.e. headings) how you wish, but you must completed all the tasks above, and in the specified order. Remember to justify any modelling choices you make. Imagine you are producing this report for a third party, for instance, as a recommendation to practitioners in the civil service. You **should not place any direct R output in the report**, except for **figures where included must be referenced in the text**. *You are allowed one page of Appendices to attach additional figures if you desire.*

- **The final report should be in pdf form and submitted as: *"student_id_report.pdf"***
- Upload the report to the coursework submission on Moodle

R-Code

In addition to the report, you should submit a **commented R script** which performs all the analysis in your report. This must be a working (i.e. running) script and libraries must be included (library(xyz)) at the start. ***Remember, I will have to run these scripts with no access to data/R-objects in your local-environment, stick to libraries discussed in the course.***

- Remember, in order to enable reproducible research **always set a seed where required**.
- **Note: you must also include code to produce the predictions in the above script/-folder.**
- Your predictions should be submitted as previously discussed as: ***"student_id_predictions.csv"***
- When you submit the R script, place this with all required files in a folder, i.e. data, saved objects which you load, and predictions. **Call this folder: "student_id_code".** Make sure your code runs from the directory before submitting!
- Compress (zip) the folder so it is called **"student_id_code.zip"**
- Upload to the courswork submission on Moodle.

Appendix A

Here is a description of variables contained in 'DM-reconv.sav'

OINUM Offenders Index identification number **Nominal** *(Do not use this variable in your analyses)*

SEX Gender of offender **Nominal**

1 Male

2 Female

TARGAGE Target age: age at first conviction in 1990 **Ordinal**

1 21/22

2 26/27

3 31/32

4 36/37

The pattern of possible ages at the time of conviction in 1990 is due to the five year gaps in the birth years due to the way the data was collated. Offenders may or may not have had their birthday by the target date.

CUST - Whether or not there was a custodial sentence at target conviction **Nominal**

0 Non-custodial sentence

1 Custodial sentence

YOUTHCUS - Number of previous youth custody sentences **Continuous**

NUMCONV - Number of convictions prior to target conviction **Continuous**

LENPRECC Time (years) from start of convictions to target conviction (this is the "criminal career" length up to target conviction) **Continuous**

TARGOFF Principal (main) offence at target conviction **Nominal**

1 Violence against the person

2 Sexual offences

3 Burglary

4 Robbery

5 Theft

6 Fraud and forgery

7 Criminal damage

8 Drugs offences

9 Other (non-motoring) standard list offences

10 Motoring offences

11 Unknown/non-standard list offences

The following 11 variables indicate whether or not the offender has been convicted for particular types of offence at any time up to and including the target conviction (they correspond to the 11 subtypes in TARGOFF):

VIOL Offender has previous conviction(s) for violence

SEXOFF Offender has previous conviction(s) for sex offences

BURGLARY Offender has previous conviction(s) for burglary

ROBBERY Offender has previous conviction(s) for robbery

THEFT Offender has previous conviction(s) for theft

FR_FORG Offender has previous conviction(s) for fraud and forgery

CRIMDAM Offender has previous conviction(s) for criminal damage

DRUGS Offender has previous conviction(s) for drugs offences

OTHER Offender has previous conviction(s) for 'other' offences

MOTORING Offender has previous conviction(s) for motoring offences

DK_NONSL Offender has previous conviction(s) for 'unknown/non-standard list'

Nominal

1 Yes

0 No

The target or outcome variables is:

RECONV Offender was reconvicted by the end of 1999

Nominal

1 Yes

0 No

Appendix B - Marking Scheme

The assignment of marks will be split primarily between the ability to perform model estimation via R, and the quality of the report. A small amount of marks will be allocated to the predictions.

- **Report, Interpretation, and Exploratory Data Analysis:** 60% - General quality of exposition, structure of report, accurate interpretation and description of methods. Where a statistical modelling choice is made I will look for a good motivation of this choice. A good report will include (and relate) to references in prior work. *Reports which are more than too long will penalised!*
- **R Code + Predictions:** 40% - Quality of commenting, not every line, but if you are using a function which requires modelling choices there should be a note, also, if you develop any functions these must be described. *The code should run without error in a new R environment, remember I do not have access to your local environment!* Your predictions must be in the format specified. If the procedure to produce predictions is not clearly outlined in your code (or is not based on your fitted models) you may not recieve any marks for this component.