DATA7202 Assignment 1 (Weight: 20%)

Due: 31 Aug 2020

Analysis of UCI online news popularity dataset: exploratory data analysis, prediction, evaluation and inference with generalized linear models.

The dataset and associated information can be found at
https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

We will focus on predicting the popularity of online articles published by Mashable over two years from 7th Jan 2013, as measured through the number of times an article was shared online.

In particular, we try to predict popularity based on explanatory variables (features) which would be available before the publication of the article, as mentioned in the accompanying paper by Fernandes *et al*. (2015) who collected the dataset. The articles have already been heavily processed to produce a large set of numerical and categorical attributes, and these are what you should work with. Many more types of attributes could be extracted from the articles and related information, but we will not attempt such processing.

We will try to predict popularity in two ways. The first target is as it is recorded, as the number of times the articles is shared over the period. The second target is a binary variable which discretises the above. In Fernandes *et al*. (section 3.1), an article is considered popular if it exceeds 1400 shares. Here we will reduce that threshold slightly and define an article as popular if it is shared 1000 or more times. You should create this binary variable. You should use multiple linear regression for the first task and binary logistic regression for the second. No model or variable selection is to be performed as part of this assignment (except as specifically mentioned in questions), despite the reasonable temptation. However, transformation of variables is allowed and variables can be removed to help compliance with assumptions or due to other problems with them.

The aim of the assignment is to develop a solid theoretical and practical understanding of the models, rather than to achieve the best possible performance. Even within the constraints above, a large number of models are possible, and the Fernandes *et al*. article indicates some of the many other options. I would also like you to consider the aims of the study. These are somewhat vague, but two aspects mentioned by Fernandes *et al*. (p536) are:

(i) “Predicting such popularity is valuable for authors, content providers, advertisers and even activists/politicians (e.g., to understand or influence public opinion).”

Why is not stated, but it seems likely that predicted popularity or probability of being popular could be considered when deciding whether or not to publish a new article, for example one submitted by a freelance journalist.

(ii) “allowing (as performed in this work) to improve content prior to publication”.

This considers how to potentially improve the (predicted) appeal of an article by considering the effect of various changes to the article, and is discussed in section 3.2 of Fernandes *et al*.

1. (i) Give relevant equations and assumptions to describe the general linear model and the logistic regression model for multivariate data (assume X has dimension p and Y is a single variable in each case). Define any notation used.

(ii) Briefly describe the algorithms used to fit each of these models and their mathematical basis.

(iii) Derive the maximum likelihood estimate of σ^2 for multiple linear regression (= the general linear model). List any assumptions. Note: consider the matrix form of the model and consider using some of the identities found in the Matrix Cookbook, available for example at https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf .

(iv) In lectures, interactions have been described between two continuous variables and between two categorical variables. Explain the interaction between a continuous variable and a binary categorical variable. Give an example, including equations and a plot. Also attempt to find evidence of one such interaction in this dataset - report your evidence and conclusions.

2.

(i) Perform exploratory data analysis as relevant to the construction of the regression models. Investigate and highlight any apparent structure in the data.

(ii) Search for at least one reasonable use of an interaction term between two explanatory variables and include it in your model. Note that the variables can be of any type, but the interaction must be in addition to the one considered in 1(iv) and you must justify why you considered this interaction (relevant graphs or tables) and explain the effect of including it in the model.

(iii) Carefully evaluate whether or not the assumptions of a multiple regression model hold. Consider, provide evidence, act to remedy (if reasonable) and discuss: standard assumptions of multiple linear regression, collinearity, any outliers or influential observations. Note: you may have to consider the conditions many times as you adjust the model (including for transformations, as discussed below).

(iv) Attempt to choose and use at least some transformations of some of the variables to improve model fit and compliance with assumptions. Justify these transformations by
(a) investigating the increase of R^2 and
(b) the assumptions met or slightly improved after the transformation comparing to 2(iii)

3. Use a general linear model (multiple regression) to attempt to predict the number of times an article will be shared based on the other variables which are available before publication.

(i) List and explain which of the variables you think would and would not be available before publication.

(ii) Give a table including all the estimated model parameters (including error variance), confidence intervals, test statistics and p-values.

(iii) Interpret the two most significant slope parameters.

4. Use a logistic regression model to attempt to predict whether or not an article will be popular (defined here as ≥1000 shares).

(i) You should attempt to check all assumptions of the model and report on this.

(ii) By default, re-use any transformations or other processing of the X variables attempted with multiple linear regression. Comment on whether any further transformations or re-representations of the data may be useful. Use them if you think they help.

(iii) Give a table including all the estimated model parameters, confidence intervals, test statistics and p-values.

(iv) Interpret the two most significant slope parameters.

5. Evaluate the predictive performance of each of the above two models using two appropriate metrics. Include details of each metric and its advantages and disadvantages. Compare your results with logistic regression to the reported results of Fernandes *et al*. with random forests.

6. For the aims of the study, as first outlined by Fernandez, and any related aims which you might reasonably ascribe to Mashable, which of the two models (multiple linear regression or logistic regression) do you think is more useful and why? This is mainly an argument about whether or not it is worthwhile to discretise the shares variable.

7. The model used by Fernandes *et al*. was a random forest. You may not know this model, but it is essentially a large ensemble of decision tree models, with generally strong predictive capabilities, but weak interpretability. It is hoped that the regression models you used will be more interpretable.

(i) Explain how one can determine how to improve the predicted popularity of an article with either of the regression models considered here.

(ii) Using your two fitted regression models, identify the articles with the highest predicted popularity and predicted probability of being popular.

(iii) For each of your two fitted regression models, list the attributes of two hypothetical articles (fake news?) which would give the highest possible predicted popularity and predicted probability of being popular, respectively. You should give values for every attribute, but for each variable, keep them within the range seen in the dataset.

(iv) Based on your analysis of dependence between variables in the dataset, comment on whether or not these hypothetical articles could be produced.

Notes:

Where possible, give reasons for your answers. Avoid thinking "you know what I mean". Say what you mean and don't assume much.

You will need to submit two files:

1) A report (pdf format) answering the questions given here.

2) Code in a zip file.

Name your files something like Yourfirstname_Yoursurname_DATA7202_assn1.pdf to help avoid confusion in marking.

You should de-emphasise any R commands and raw output in the report. R commands should go in the code file. You should not be including all R output. Any output should be edited or re-formatted to reasonably present the most important information. Use figures and tables with numbers and captions for each and refer to them from the text. Try to make sure that any figures you produce can be fairly easily understood.

As per http://www.uq.edu.au/myadvisor/academic-integrity-and-plagiarism, what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. Equations are either correct or not, but you should use consistent notation throughout your assignment, define all of it and ensure that your report flows logically.

You are asked to use the R software environment for this assignment. This is available on all computers in the Maths Department and is also free to install on any of your own computers. Information and downloads are available from http://www.r-project.org/ . Rstudio https://www.rstudio.com/ is a quality free interface for R.

Submit your pdf report via TurnItIn on Blackboard. Please make sure that the overall file size is less than 10 MB. It is not necessary to include every data point on a typical plot. Using just a sample of the data can be ok for most illustrative plots. The main exception is if you want to show outliers along with all the data.

References:

A.J. Dobson, and A.G. Barnett, *An Introduction to Generalized Linear Models*, 4th edition, CRC Press, 2018.

J. Faraway, *Linear Models with R*, 2nd edition, Chapman & Hall/CRC, 2014.

J. Faraway, *Extending the Linear Model with R*, Chapman & Hall/CRC, 2016.

K. Fernandes, P. Vinagre, and P. Cortez, *A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News*, in: Pereira F., Machado P., Costa E., Cardoso A. (eds.) Progress in Artificial Intelligence, EPIA 2015, Lecture Notes in Computer Science, vol. 9273, Springer.

J. Maindonald and J. Braun, *Data Analysis and Graphics Using R - An Example-Based Approach*, 3rd edition, Cambridge University Press, 2010.

W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th edition, Springer, 2002.

S. Weisberg, *Applied Linear Regression*, 3rd edition, Wiley, 2005.

H. Wickham, and G. Grolemund, *R for Data Science*, O'Reilly, 2017.