STA304XS – Intro to Machine Learning.

Assignment 2 2024

September 27, 2024

Question 1 (25 marks)

You are a Data Scientist working at Twitter, the social media platform where users can make public text, picture, and video posts, and where other users may engage with said posts by commenting, retweeting, or liking posts. Since the viability of the platform fundamentally depends on how frequently users post tweets, the Data Science team is interested in which factors are predictive of the number of tweets an individual will make. You are given a number of observations on the following variables:

Variable	Description
`Tweets`	**A count** response, counting the number of tweets an individual made during the period of the study – 24 hours.
`Connectedness`	Continuous: A normalized measure of user connectedness on the platform. (Graph-based statistic.)
`Follower_Engagement`	Continuous: A normalized measure of follower engagement with the user.
`Sentiment`	Sentiment is a categorical variable measuring overall sentiment of an individual's posts and takes on three values, negative, neutral, or positive.

The Data Science team wishes to analyse relationships between the predictors and the response. As part of your analysis, you are tasked with writing various bits of code and subsequently making interpretive assessments through the questions below.

Note: I've added leading indicators to each question telling you what should be reflected in your compiled write-up. To that end, here's what they mean explicitly:

- R code: This means, I want to see your R code typeset at that point in your document. So make sure the code-chunk options are set such that the code renders.
- Plots: This means, I want to see a plot rendered in the write up.
- In-text response: This means I want a paragraph-style response where you type out your response in one or a few sentences.
- Combinations of the above: If you see R code + plots, that means render the R code and the plot, and no typed response is required. If you see R code, that means just render the R code and no other response is required.

Question 2 (5 marks)

Exploratory data analysis: The first step in most statistical reports is conducting exploratory data analysis. This means investigating data, calculating basic summary statistics, and or non-parametric summaries of aspects of the data. A model-free investigation of the data.

(a) R code + plots: (3)

- Draw appropriate scatter plots of the response and the continuous predictors.
- Draw boxplots for any categorical predictors.

(b) In-text response: Based on the plots in (a), interpret the empirical relationships between the response and predictors. (2)

Question 3 (8 marks)

Modelling: Now you build statistical models to probe relationships in the data. And once you have vetted the models, you can draw conclusions about the data from the relationships predicted under the models.

(a) R code: Fit a tree-model to the data using the `rpart` library. Set the stopping criterion `cp = 0.04` and set the minimum number of observations required to affect a split to 4. Run `set.seed(1)` before fitting the model. (1)

(b) R code + plots: Construct a validation plot and add a vertical line to the plot indicating your choice for the level of pruning to apply. Also, plot the decision tree that results from pruning. Include these plots in your write-up. Hint: `abline(v = ...)`. (2)

(c) Typed response: Are any of the variables/heuristics redundant for purposes of predicting the number of tweets? Clearly motivate your response. (1)

(d) R-code + plot + written response: Use the `h2o` package to fit a $(5, 5)$-network to the data, using a learning rate of 0.01, 1000 epochs, and L2-regularisation. In order to determine the level of regularization to apply, split the data into a training and validation set $(80 : 20)$, use `set.seed(2024)` before splitting the data), and fit the network for various levels of the L2 regularisation parameter. Measure the validation error for each value of the regularisation parameter and save that in a vector. Plot these values as a function of the regularisation parameter and report the relevant level of regularisation to be applied to this model. Also explain why conducting a validation analysis is relevant for determining the existence and nature of relationships between the predictors and responses here. Hint: Set the seed in the `h2o.deeplearning` function to 2. Hint: `exp(seq(-10,-3, length = 20))`. Further clarification: Don't set `distribution = 'poisson'` despite what I said in the video. We're extracting the error via `h2o.mse` and it is not clear that this is the objective trained on – I suspect it is not. But for the sake of this exercise and an exam, if I give you a regression problem just use quadratic. The implied property being modelled is the same here so won't matter much. (4)

Question 4 (6 marks)

Response curves:

(a) Typed response: For a tree-based model, which model do we use to construct a response curve? The pruned model or un-pruned? Clearly motivate your response. (1)

(b) R code + plots: Construct and plot a response curve **over** `Follower_Engagement` and `Connectedness` **for** an individual with neutral `Sentiment`. Do this for both the neural network and tree model and include the response curves in your write-up. Further clarification: Everything that comes between **over** and **for** must vary **simultaneously**. Everything after **for** is fixed. (3)

(c) Typed response: Interpret the response curves. (2)

Question 5 (6 marks)

Further Technical Analysis

An alternative means for conducting the validation analysis using a **tree-based model** (NOT a Neural Network) is as follows: Split the data into a training and validation set by randomly assigning 80% of the observations to a training data frame and the remaining to a validation data frame. Then, fit the tree model to the training data frame for a sequence of `cp` values and predict responses for the validation frame for each value of the `cp` parameter. Then, measure the model fit by way of the negative of the log-likelihood under a Poisson model (since the responses here are counts). We can then construct a validation curve over varying levels of the complexity parameter and choose the most appropriate value. For these purposes,

$$\text{minus-log-likelihood}(cp) = -\sum_{i=1}^{N}\left[y_i\log(\hat{y}_i^\star(cp)) - \hat{y}_i^\star(cp) - \log(y_i!)\right]$$

where y_i is the observed count response and $\hat{y}_i^\star(cp)$ is the predicted response under a tree fitted using `cp`. Use this expression and the questions that follow to conduct the validation analysis.

(a) R code: Manually create a 80-20 validation split of the data. Use `set.seed(2)` before running the analysis. (1)
Further clarification: Manually just means like I did in class with the `sample` function.

(b) R code + plot: Conduct a validation analysis using a sequence of 100 `cp` values ranging from e^{-8} to e^{-4} and plot the negative log-likelihood as a function of this sequence. (Set the minimum number of observations required to affect a split of 4 as before.) (3)
Hints: This is similar to how one would do a validation analysis for neural networks; `predict(tree_model, data_val)`; Further clarification: leave the NAs as is. You don't have to do special treatment of the zero-predictions. You also don't have to set anything to Poisson in the rpart function.

(c) Typed response: Would the analysis of questions 3 and 4 change under this validation mechanism? That is, does the model fitted using the `cp` parameter selected here differ from what was selected before? Clearly motivate your response! (2)

Question 6 (12 marks)

You are a Data Scientist consulting with an online retailer. The retailer sources various wines from the Western Cape region and brands the bottles before storage and sale. The marketing strategy is as follows: The 'brand' is called 'Rand(U)', a nod to random number generators where the premise of the brand is that a wine of random source is contained in a bottle and wine types are only distinguished by the (seed) number assigned to U. The retailer wishes to distinguish wines using only a few distinct numbers. Your task is thus to find natural distinct groupings of the wines, whatever their source might be. You are given 178 observations on the following measurements:

Variable	Description
`Alcohol`	Scaled alcohol content.
`Color_Intensity`	Scaled score of colour intensity liquid
`Hue`	Scaled score of hue of the liquid.

(a) Write an R-function `k_means(X,K,iter)` that applies Lloyd's algorithm for `K` clusters and `iter` iterations. This function should return a list with `D`, the distance from each observation to the `K` cluster centroids, `SSQ` the within-cluster-sum of squares at the final iteration, and `z` the cluster index for each observation. I.e., the cluster it belongs to. Clarification: in the R code in class, I used `initial_indices` or something like that where `K` is here. I actually meant that `K` here should be an integer (just the number of clusters). So, in your function you should sample `K` indices from `N` without replacement. So, something like: (2)

```
ind = sample(1:N,K,relace = FALSE)
Mus = X[ind,]
```

Further clarification: Use `set.seed(2024)` just before running the analysis so we all get the same results.

(b) Write an R-function `SSC(res_k_means)` that takes the output from `k_means(X,K,iter` and calculates the **simplified silhouette value** for each observation, the **average simplified silhouette score** for each cluster, and the **overall average simplified silhouette score**. This function should return a list of objects `s`, `s_bar_k`, `s_bar_overall` with the relevant quantities. Typeset your function to display as the response to this question – an R chunk displaying your function. Further clarification: Don't use the built-in function for this! That doesn't calculate SIMPLIFIED s-scores. I want you to do this manually! (4)

(c) Calculate and plot the within-cluster sum of squares and overall average simplified silhouette scores for $K = 1, 2, \ldots, 6$ and interpret the results. (Use these plots to give an appropriate number of groupings of wines.) Clarification: When I'm asking you to interpret the plots here, I'm asking you to pick a number for K, the number of groupings in the data based on the plots. Also, what does that mean in terms of the problem? (What should the retailer do with this info?) (3)

(d) Are the groupings well-separated? Use the relevant statistical tools to motivate your response. Clarification: Here, I'm asking about the extent to which distinct groupings can be found. Use the statistics (values and behaviour of the statistics) in (b) and the plots in (d) to motivate your response. (3)

Statistical Report Writing Conventions

- You may use any typesetting software to compile your report. Rmarkdown and LaTeXare preferred for the obvious reasons, but you are welcome to use whatever you are comfortable with as long as your final hand in is a legible **PDF** file.

- The main body of your write-up **may not exceed 9 pages**.

- Include your code in an appendix. (This will not count to the main body of the report.)

- **Do not include figures in an appendix.** Figures are supplemental to your writing and should be included in body of the write-up. Also, figures presented on their own are rarely of any value. The only species of figure that can live on its own in this context is an infographic. Figures on the other hand are graphical mechanisms which support discussion in scientific reports. Don't conflate their use in report writing for the standalone figures you find in slides/handouts: In the absence of an author, text, or context, standalone figures present as disembodied thoughts left on the page for the reader to step in and then clean off their shoes. Be nice to your readers :).

- Include a plagiarism declaration as the very last page in your report. No signed declaration, no mark. I've included an example Rmarkdown file showing how you can incorporate a pdf directly in your markdown compilation. Include one for each student who worked on the project.

- Do not simply copy console code or screenshot and paste in the body of your write up unless specified otherwise. You are the analyst, not the reader. A well written statistical report would not contain any console output. Tabulate and typeset or plot your output properly. (I've included an example of how to tabulate R objects in the Rmarkdown file.)

- Use the naming convention `STDNUM001_STA304XS_A2.pdf` for your file. Note the underscores. Use the student number of the person who submits **AND MAKE SURE ALL GROUP MEMBERS' names are INSIDE THE DOCUMENT.**

- A single file with all of your R code must be uploaded separately to the code tab for the assignment. Use the naming convention `STDNUM001_STA304XS_A2.R` for your file. Note the underscores. Your R code should NOT contain any of the following:

  ```
  install.packages()
  rm()
  setwd()
  ```

  We want to be able to run your code on my computer without having to manually edit your code, installing libraries or calling to external files.

- If you are working in a group, use a single student number for the naming convention. It should be that of the person handing in. The names of your collaborators should be included in the document. You may work in groups of up to three. (1,2, or 3 students per project.)

Note that ANY deviation from the above naming, code, and page limit conventions WILL be penalized. I've also reserved 3 marks for the typographical quality of the report.