Internet Advertisements: A Classification problem

Rushil Sheth

May 1st 2017

Introduction

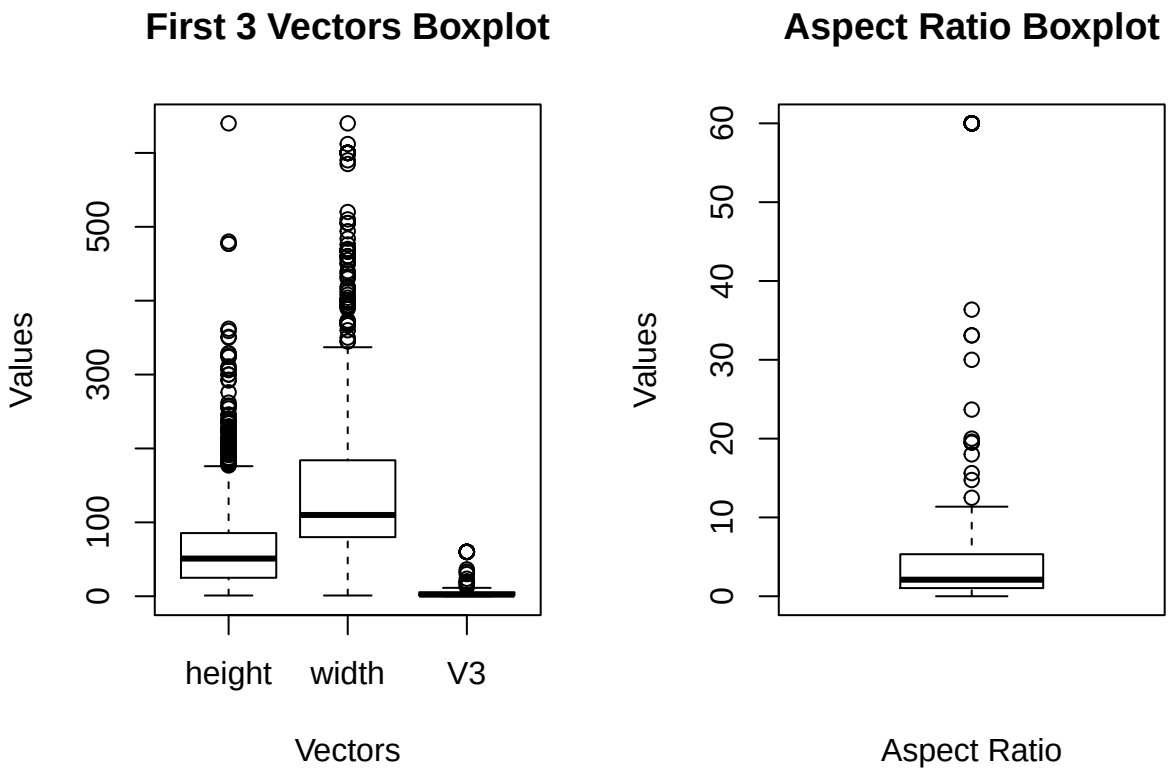
In this project we will attempt to build a classifier, which determines if an image on a webpage is an advertisement or not. Our final model is correct over 95% of the time. Our methodology and reasoning, ending with a discussion will be given below.

Data Description

The data set we are given is quite large, with 3,279 rows and 1,559 columns. Specifically, the data contains 3,279 different images along with 1,558 corresponding attributes which help explain whether an image is an advertisement or not and the last column is "ad." or "nonad." There are 2821 non-advertisements and 458 advertisements.

The 1,558 potential predictors are the height, width and aspect ratio of the image and the rest are variables that take on the value 0 or 1. We have no idea what the meaning of the individual variables are except that they help predict whether an image is an advertisement or not. Some of these binary variables are drawn from things such as: text phrases in URL for the image, URL for the page, and other info about the webpage, to name a few.

Nearly one third of the images have missing values for the height, width and aspect ratio variables. To alleviate this problem we will enumerate the missing values of these variables using the median. We will also add some random noise to the medians, so that we avoid shifiting or skewing our data.



From the box plot on the left, we can clearly see that for vector V1 and V2 we should use the median rather than the mean, because the median is a much better measure of centrality here, given all the outliers for their data.

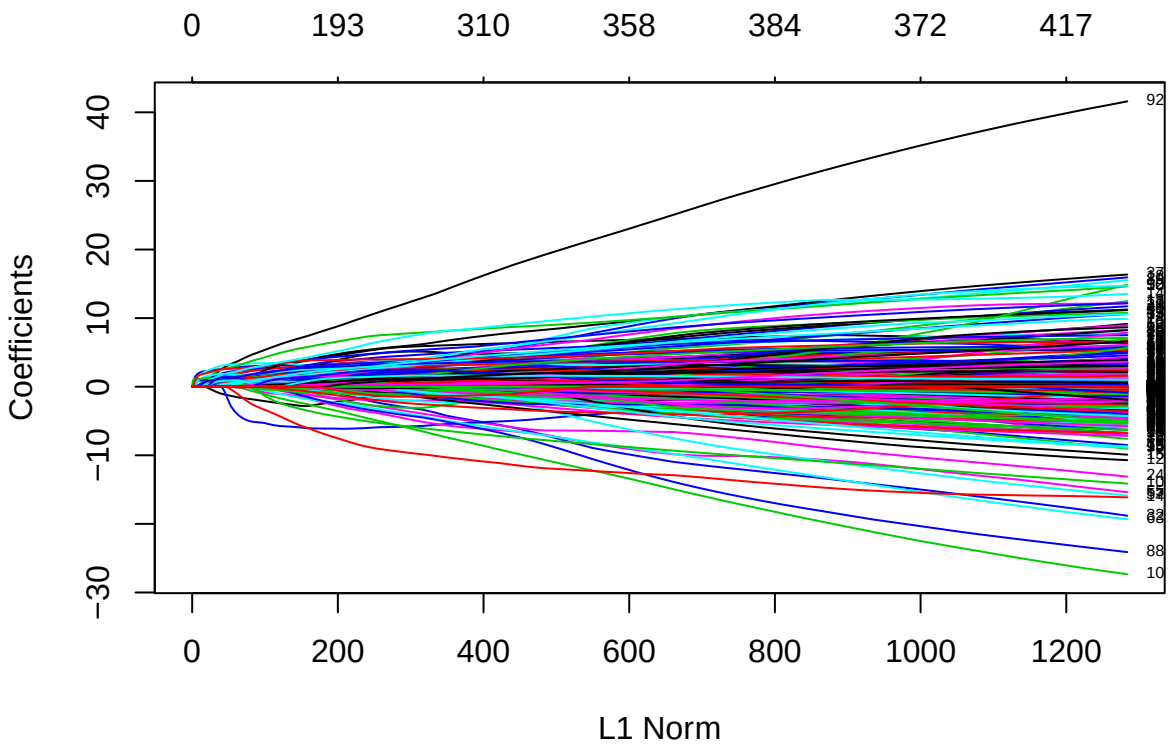
A better look at vector 3, which describes the aspect ratio, using the boxplot on the right, we can see that the median is also a much better measure of central tendency, for similar reasons as vector 1 and 2.

Methodology and Results

Due to the extremely high amount of predictors, I decided to use a shrinkage method, specifically lasso logistic regression. Lasso is a penalized linear regression, which drives many of the predictors(irrelevant ones) coefficients to zero. I use lasso **logistic** regression because this is a classification problem.

We split our data into a training and a test set. Our training set contains a random(seed is set to 3 for reproducibility) 80% of the data and the test set contains the other 20% of our data. We find the ideal lambda using cross validation on our training set and the `cv.glmnet()` function from the package, `glmnet`.

We can also see from the plot that there are a few coefficients, namely 92, 10 and 88, which are much higher than others, showing the effectiveness of lasso.



After building our model using the training set, we now use the `predict()` function and set our threshold, which determines if an observation is an ad, as .5. We choose .5 because of the majority rule used in logistic regression. We see how well our model does by calculating misclassifcation error on our test set:

$$\text{misclassifcation error} = 1 - \frac{\text{number of correct prediction}}{\text{test set size}}$$

We have a very low misclassification error. We can interpret it as when using our model to predict whether an image on a website is an advertisement or not we will only be wrong about 3% of the time. This is a great and very satisfactory result.

	Lasso Misclassification Error
1	0.0335

Discussion

Now we are tempted to use a different type of penalized logistic regression. Here we use ridge logistic regression. Ridge is a shrinkage method, which addresses the problem of correlated predictors. Ridge, however, still allows for the problem of overfitting.

Our misclassifcation error here is also extremely small and is very satisfactory. We can compare the two misclassification errors, but remember that these can change due to randomness and test/train sets choosen.

	Method	Misclassification.Error
1	Lasso Logistic Regression	0.0335
2	Ridge Logistic Regression	0.0381

Finally, this model may **not** be generalizable to a larger data set of the same structure, since there are significantly more non-advertisements than advertisements in our given data set. This potentially skewed our model, but overall given our data set, we should be very satisfied with our model.

References

R packages:

- DataComputing
- glmnet
- xtable