Student Data Set

Álvaro Bermejo García

February 2016

Data Set

Since becoming a democracy Portugal has been fighting an uphill battle, get on par with the rest of Europe.

This battle had many fronts, from economic, to social, political, etc...

As of today Portugal has achieved most of what's expected of an European country, but there is still one aspect where Portugal is seriously lagging behind, **Education**.

Right now Portugal has a 40% school drop out rate, compared to 15% European average, for the last years the country has expended a sizable portion of the budget in trying to understand why and how to fix this problem.

And that's where this research comes from, it is not really trying to understand the causes and where efforts could be made, but studying the viability of machine learning strategies in this kind of real world situations.

Input Details

Data comes in csv files, accompanied with a `R` script that can be used to merge data from both subjects, unfortunately when exporting that data structure from `R` to `Octave` it is a collection of hashes, not a regular matrix, which makes it harder to work with it.

Because of that and because the number of students from portuguese the subject (649) was much greater than the number of math (395) we only included those results from the portuguese subjects here, although in measures with the math data set results were very close (around ±3% variance).

Because the csv file used characters to represent data we had to convert everything to numbers, that procedure can be seen in `getData`.

```
function [X,y] = getData(filepath)
    preProcessData(filepath);

    data = csvread([filepath '.oct']);

    X = data(:, 1:32);
    y = data(:, 33);
end

function preProcessData(filepath)
    filestr = fileread(filepath);
    filestr = substr(filestr, 229);
    filestr = strrep(filestr, 'GP', '0');
    filestr = strrep(filestr, 'MS', '1');
    filestr = strrep(filestr, 'M', '0');
    filestr = strrep(filestr, 'F', '1');
    [...]
    filestr = strrep(filestr, '"', '');
    octcsv = fopen([filepath '.oct'], 'w');
    fprintf(octcsv, '%s', filestr);
end
```

Columns 1 through 30 are related to socio-economical status like parents situation, alcohol consumption, whether they are on a relationship, etc..., 31 and 32 are first and second semester marks.

Logistic Regression

```
λ = [0,0.1,0.2,0.5,1,2,5,10,15,20,25];
```

Using `fminunc` with a max of 1000 iterations.

Neural Networks

```
λ = [0,1,10,100];
hidden_nodes = [0,1,10,100];
```

Using `fmingc` with up to 1000 iterations were allowed, but the best result hover over ~96 iterations.

Support Vector Machines

```
C = [0.01,1,10,100];
σ = [0.01,1,10,100];
```

A Gaussian kernel was used, as it was the one which had the best results.

Results

Percentages

This table reflects the best results achieved, changing not only, hyper-parameters but also the confidence threshold, these can be seen on both figures 1 and 5.

In bold the best percentage of each category.

Method	With previous marks	W/o previous marks
LR	85.28%	**69.94%**
NN	**88.72%**	69.74%
SVM	83.97%	66.41%

Graphics

Following some graphs showing the breakdown of accuracy, recall, learning curves or the adjustment planes of the different methods

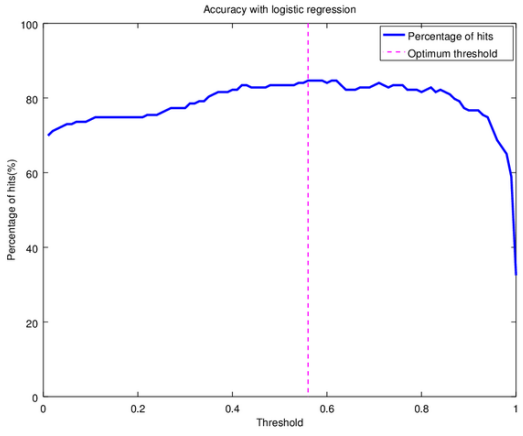

Figure 1: Logistic Regression Accuracy

Logistic regression Accuracy is more or less stable even when changing the threshold.

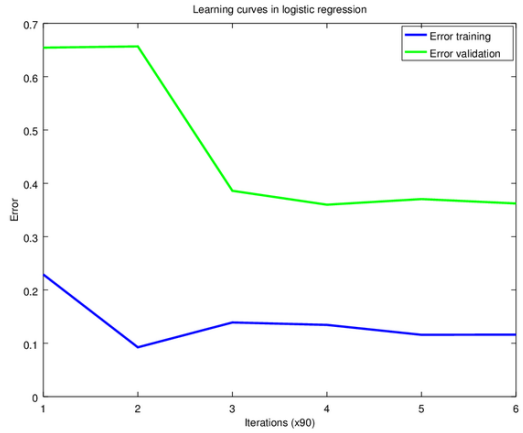

Figure 2: Logistic Regression Learning

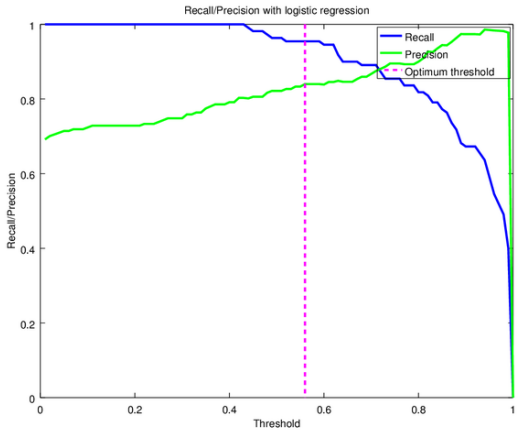

Figure 3: Logistic Regression Recall

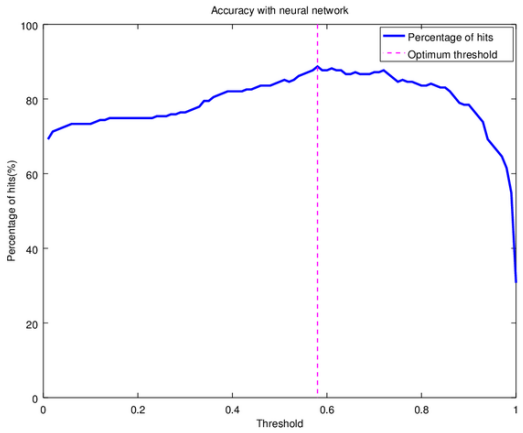

Figure 5: Neural Network Accuracy

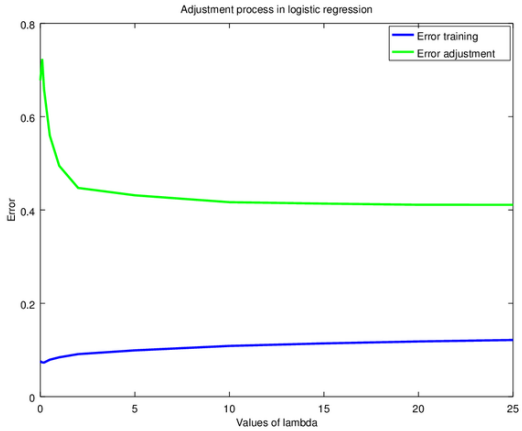

Figure 4: Logistic Regression Adjustment

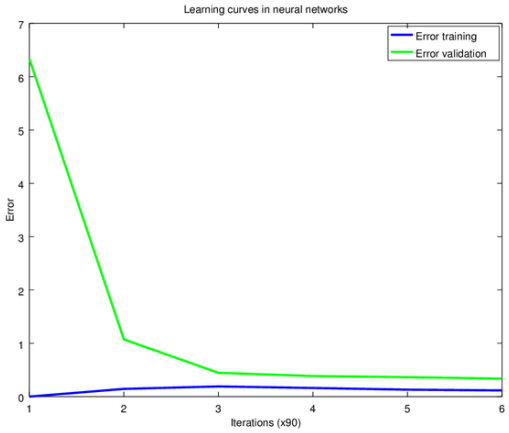

Figure 6: Neural Network Learning

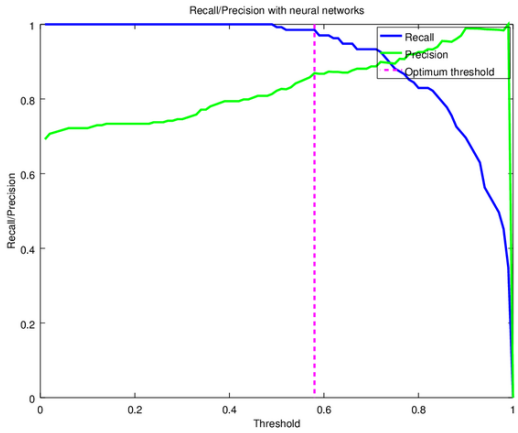

Figure 7: Neural Network Recall

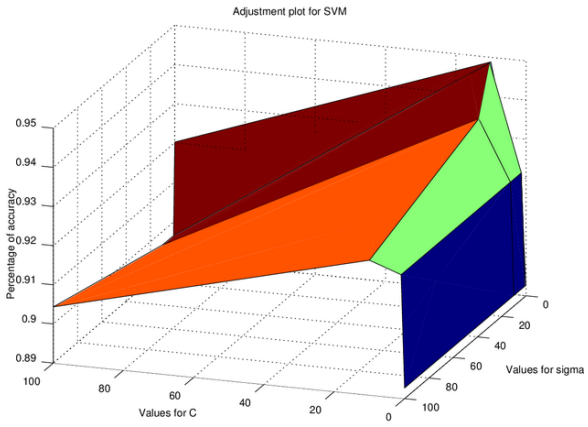

Figure 8: Support Vector Machine Adjustment

Possible Improvements

K-Fold Cross Validation

`Matlab` `crossvalind` functions makes this pretty easy, unfortunately `Octave` has no direct equivalent, and although it is not exceedingly hard to code it, it is not trivial to make it generic enough so it works for every size of data.

There is a half-coded implementation of a 10-fold validation but it tended to produce some errors with certain example sizes.

Parallelization

While the Neural Network in the code has parallel capacities because `fmincg` (the gradient descent algorithm we use) seems smart enough to take advantage of multiple cores the SVM and the LR did not had any parallel capacity, this meant that on our 4-core machine we were wasting a lot of cycles.

To illustrate this even with this small set of data running the three strategies could take upwards of half an hour, this is including the adjusting option on all three, even then, seems like very poor performance with no scalability.

This could be trivialy parallelized with `Matlab` `parfor` syntax but in `Octave` this is syntactic sugar for just a `for` loop, the only analogue option is the `parallel` package in the Octave forge, unfortunately there wasn't just enough time.

Better Stratification

The stratification used this time was a bit *manual* in the sense that in order to mix the slices a variable had to be be changed on the source code, it would have been great to have automatic random stratification, but unfortunately, again, not enough time.

Decision Trees

On the original paper the authors get a much better result with decision trees and naive Bayes tha with plain machine learning algorithms, but because we set out to not use any library but what was coded by us we didn't had any implementation of decision trees at the time

Conclusions

Predicting Student Performance only with socio-economical context is hard, however with previous academic results we can not only provide an accurate binary assertion on whether the student will pass or fail, but we can with relative precision the range where the mark will be.

These results seems to indicate that with enough data machine learning could be successfully applied to education analysis.

Another point to make is that even-though Machine Learning Techniques are not the most adequate to try to examine and extract what attributes where taken into account the most we can examine the theta vector given by the logistic regression to infer the following:

- Trying to analyze & predict anything to do with human behavior is hard
- When we don't take into account the previous mark the absences and the desire to study higher education are the most important factors
  - Age plays quite an important factor (The lower, the better)
  - Doing extra activities is as important as going to paid extra classes
  - Previous school years failures are also important (And related to age)

References

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Code & Data