Project 2: Classification and Regression

from linear and logistic regression to neural networks

Anders Eriksen

Contents

The main summary of the work

1 introduction

aims and rationale of the physics case, what you've done as well as a brief summary of the report
o Motivate the reader
o What done
o Structure of report

The project aims to reproduce the results of the artificial neural network predictions for credit card default studied by Yeh and Lien [2009].

2 Methods

Theoretical models and technicalities.
o describe methods and algorithms

o explain implementation of methods and say something about the structure of algorithm and present parts of code
o Plug in some calculations to test code, such as selected runs used to validate and verify results. Latter extremely important. Reader needs to know that your code reproduces selected benchmarks and reproduces previous results, either numerical and/or well-known closed-form expressions.

The main aim here is to study both classification and regression problems. With an intention to reuse the regression algorithms studied in project 1 of the course, Eriksen. Further we include logistic regression for classification problems. Additinally implementing multilayer perception code for both regression and classification problems. Mean squared error, cross validation algorithm as well as the R2 or the accuracy score for classification can be utilized here.

The data sets used are terrain data from the previous project as well as UCI's credit card data set. These are for the regression and classification respectively.

The first point of order is to familiarize ourselves with the data. Import and organize the set to gain some idea about the shape, size and layout of the data. Following this comes defining the cost function and design matrix. These should be mostly the same as during the first project, Eriksen. A gradient descent method would also likely be helpfull here, as well as further on in the project. This method could also be compared to scikitlearn's modules.

The Credit card dataset consists of roughly 3000 people surveyed with each having 24 different categories such as marital status, sex, previous payment details and so on. The data set is gathered from a *.xls* file, and organized into a pan data structure to organize. The design matrix here, is the set, sans the output in the column of whether or not the person defaulted their credit.

In order to measure the preformance in the classification problem a so-called *Accuracy score* is used. This is the number of correct guessed targets t_i over the total number of targets. A perfect classifier would yield an accuracy of 1.

$$Accuracy = \frac{\sum_{i_1}^{n} I(t_i = y)}{n}. \tag{1}$$

I here is an indicator function returning 1 if $t_i = y$ and 0 otherwise in this instance of a binary classification. y_i is the output of the Logistic Regression code. The accuracy can be compared with the scikitlearn for performance benchmarking.

Once the groundwork is laid, we can begin work on a feed forward neural network. The back propagation algorithm based on course slides, Hjorth-Jensen, can be seen bellow in 1. A discussion of cost function is also warranted.

Algorithm 1 back propagation algorithm

```
for  $i = 0$ to $i = n$  do
    Iterate
end for
```

For the regression fit, the cost function has to be modified appropriately.

Once the calculations have been completed, the methods need to be compared and discussed to find the best result in the cases of classification and regression.

3 Results

The results and discussion of such
o Present results
o Give critical discussion of you work & place it in correct context
o Relate work to other calculations/studies
o Reader should be able to reproduce calculations should they wish to do so. All input variables should be properly explained.
o Make sure figures and tables contain enough information in their captions. Axis labels, etc. A reader should be able to get a first impression of the work by purely studying the figures and tables.

4 Conclusion

Conclusions and perspectives
o State main findings and interpretations
o Try as far as possible to present perspectives for future work.
o Try to discuss the pros and cons of the methods and possible improvements.

5 Appendix

any extra material
o Additional calculations used to validate code.
o Selected calculations. Can be listed with few comments.
o Listing of code if necessary.
Consider moving parts from methods to appendix. A webpage is also an appropriate place for a lot of this type of info.

o Always reference material you base your work on, either scientific articles/reports or books.
o Refer to articles as: name(s) of author(s), journal, volume(Bold), page and year in parenthesis.
o Refer to bookds as: name(s) of author(s), title of book, publisher, place and year, eventual page numbers.

References

Pandas: powerfull python data analysis toolkit. URL `https://github.com/pandas-dev/pandas`.

Anders Eriksen. Ordinary least squares regression of franke's function.

Morten Hjorth-Jensen. Lecture slides, mcahine learning.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.