Group Project

Cummins, E.[*], McLeod, H.[**], and Tsai, V.[***]

[*]Dept. of Information Studies
[**]INST0060 - Foundations of Machine Learning and Data Science
University College London, WC1E 6BT
{*ellen.cummins.17,hope.mcleod.14, vanessa.tsai.13*}*@ucl.ac.uk*

March 8, 2019

1 Abstract

This report describes how machine learning techniques were used to train and test classification models built using logistic regression and Fisher's Linear Discriminant. The aim was to demonstrate how models can be created with different model algorithms and parameters and how they can predict a sample's membership to one of two categories. It also demonstrates how model accuracy relies on choosing the most appropriate algorithms and modelling techniques. The ultimate goal was to produce models that are capable of predicting, as accurately as possible, unseen data.

2 Introduction

Classification is of interest because it allows automation of identification of data, whether that be to assist in decisions (e.g. medical) or predict future events (e.g. weather). Using machine learning to help categorise an object can be of great assistance to many live systems e.g. medical.

This report first gives a brief background about the algorithms used and why particular ones were chosen. Next a description of the datasets used is given with emphasis on the one chosen by the group (as opposed to the compulsory dataset given). Here some exploratory analysis will show the state of the data before machine learning algorithms are applied.

A description of the method used follows the Dataset section. The techniques applied in preparation of fitting a model and methods chosen to check the model's stability, accuracy, and the challenges encountered are presented. The python files containing the models are also mentioned.

After the Method, the Results and Discussion section will be used to outline outcomes from using th models on unseen data, the conditions in which the models worked better and what else could have been done to improve model performance.

3 Background

Logistic Regression and Fisher's Linear Discriminant algorithms were used to classify data from two datasets: a compulsory one which can be used as a benchmark i.e. the marker is already familiar with its contents and potential. The other dataset was chosen by the group to match the criteria of having 5-20 columns, 1000 instances and at least 200 instances belonging to each target category.

3.1 Logistic Regression

Logistic regression can be used for binary classification. Like linear regression, its model initially uses linear methods to output a continuous number in an intermediate step. However, unlike linear regression, it uses a threshold value to make a decision of how to map that output to discrete number from a limited set of numbers.

It is considered to be a probabilistic discriminant classifier; probabilistic because a sigmoid function is first used to calculate the probability that an instance of data belongs to a category of interest. Discriminant because unlike a generative model, this probability is directly mapped to a class using a decision step.

3.2 Fisher's Linear Discriminant

Fisher's, also used for binary classification. Fisher's does not expect data to be normally distributed or equal class covariances (remove this ref when happy with section: wikipedia LDA)

3.3 Why these methods were chosen

According to Yarnold et al, "Logistic regression analysis (LRA) and Fisher's discriminant analysis (FDA) are two of the most popular methodologies for solving classification problems involving a dichotomous class variable and two or more attributes"

Fisher's makes more assumptions about the underlying data. (https://www.stat-d.si/mz/mz1.1/pohar.pdf).It requires the independent data (inputs) to be continuous in nature (find ref) whereas logistic regression can be used to produce models from data containing independent variables which are discrete and continuous in nature.

Parameters ...

Linear Discriminant Analysis is best used when the predictors are continuous in nature (check this) therefore, the mammogram dataset might be better suited to logistic regression. https://stats.stackexchange.com/questions/130960/ to-use-linear-discriminant-function-and-when-logistic-regression. If have more than 2 classes then LDA is better but this is not an issue with these 2 datasets. Logistic regression is also better when have larger datasets. So using LDA with the abalone dataset is looking to be the better choice.

LDA assumes data is Gaussian (but FDA doesn't!) https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning

4 Datasets

Abalone Dataset (http://archive.ics.uci.edu/ml/datasets/Abalone?pagewanted=all) Abalone is a dataset consisting of 8 attributes with the 9th column representing the target variable. There are 4177 instances and no missing values.

Drugs Consumption dataset (https://archive.ics.uci.edu/ml/datasets/Drug+consumption+There are 1069 instances with label 0 and 816 instances with class label 1 which represent 'has had no drugs' and 'has had drugs' respectively.

Exploratory Analysis One of the requests for this classification task was to have at least 200 instances from each class. Once the abalone dataset has been modified so that any ring value ¿ 10 is assigned to class 1 and any less than or equal to 10 is assigned to class -1, there are 2730 instances belonging to the negative class and 1447 belonging to the positive class. This dataset is therefore going to be trained with more negative class data than positive class. This might have an effect on the outcome.

When do pair plots i.e. plot one variable against another using a scatter graph or histogram if plotting a variable against itself, can see that none of the histogram plots demonstrate good separation of the classes.

Some pair variables demonstrate a linear correlation; mainly the different types of weight variables which is no surprise. When variables show a high correlation, this can be interpreted as some variables being redundant i.e. if have variables which follow each other in behaviour, then there is no point in keeping them all. It will also mean less computation and a more efficient prediction process. Height does not show a strong correlation with weight which is a bit of a surprise.

Discrete data in abalone There is one column that is categorical 'sex'. It contains 3 values, 'male, 'female', and 'infant'. The dataset was modified by removing the 'sex' column and adding two dummy variable columns called 'male' and 'female'. For an instance, these have been encoded as 1 if male and 0 for the corresponding female column. If an instance is female, then it is encoded with 0 for male and 1 for female. If an instance is neither male or female, then both are encoded with 0. This indicates that the instance represents an infant.

4.1 Method

Logisitic Regression Logisitic regression has been chosen as it is a popular classification techique due to it being quite easily interpreted (say what mean by this). It is a useful algorithm because it's attributes can be contain a mixture of continuous and discrete data. Logisitic regression data does not need to be standardised (why?). It is therefore a good technique to be used on both datasets.

The code for logistic regression is in classification_model_using_logisitic_regression.py With regards to the abalone dataset, the script first does some data preparation i.e. assigns class values to the target column and adds 2 new dummy variables to deal with the categorical column 'sex' which contained more than 2 different values. The modified data is imported and some pair plots drawn to get an initial idea of what the data looks like e.g are there any pair of variables which split the data into classes well.

Next the data is split into train/test with 60/40 split. It was an initial decision not based on any particular logic but with the intention of potentially changing this split if a poor model results i.e. the model might be poor because there is not enough data to train with. An ROC plot is created. The ROC was used to see the model's ability to predict true positives and therefore the idea was to produce an ROC with as large an area under the curve as possible.

The plan was to improve the logistic regression model by trying various modifications: - use basis functions if the prediction is poor using a linear model. A poor linear model might be indicated from the ROC e.g. if a 45 deg plot results then the model is as likely to predict incorrectly as correctly. If a model, created using basis functions, has good predictive behaviour, then this could suggest that the data is not linearly separable. - Use regularised logistic regression if the model predicts better than with test data - try different model probability cutoff points

5 Results

Abalone Dataset

Figure 1: Abalone ROC linear

Ideally want the area under the ROC to be as large as possible. Want a consistent high true positive rate. MLE is not used to find the best weights for a model because there is no closed form solution.

When use RBF basis on the abalone training data, get ROC plot:

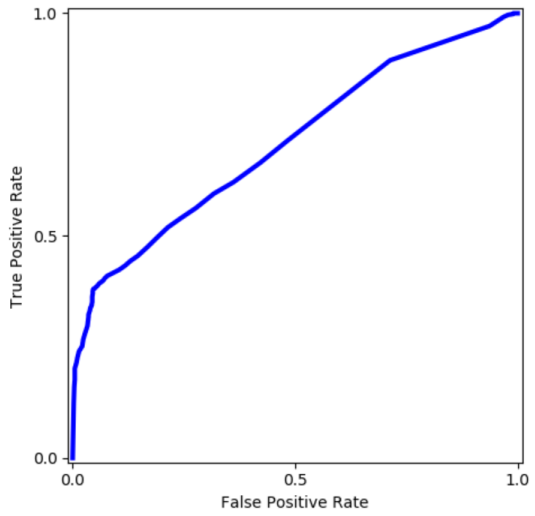

Figure 2: Abalone ROC using RBF basis

The RBF basis function was used to see if could improve the model. By using basis functions we are trying to separate the data linearly in a different feature space.

The Abalone using RBF basis on the inputs does not separate the classes well. Although not quite a 45 deg line, which would indicate a model that predicts as many true positives as false ones, it is not far off.

When the RBF is used, the 9 features are reduced to 7 features which suggests that some of the feat

6 References

Yarnold, P (1994). Optimising the Classification Performance of Logistic Regression and Fisher's Discriminant Analyses. Educational and Psychological Measurement 54(1):73-85

2004, Pohar et al: https://www.stat-d.si/mz/mz1.1/pohar.pdf