EE3-23: Coursework for Introduction to Machine Learning

Douglas Brion
Imperial College London
CID: 01052925
db1415@ic.ac.uk

1. Introduction

This report examines different approaches to predicting wine quality using the Wine Quality [1] dataset. This is a large dataset containing 4898 samples of white wines and 1599 of red. Throughout this report wine quality is examined using regression, attempting to predict the quality of a wine from various input parameters. Multiple learning methods are implemented, discussed and compared in order to obtain a predictor with the smallest test error possible.

2. Data Preparation

Both red and white wine datasets were available and it was decided to combine both into a single dataset consisting of 6497 samples in order to learn how to predict the quality of either a red or white wine.

The data provided in the now combined dataset was not normalised, therefore each attribute was normalised with respect to it's mean and standard deviation. Outliers over a threshold of $Thres = 5$ from each attribute were then removed to reduce noise in the dataset for training. This altered data was output to `winequality-fixed.csv`.

Attributes	Mean	Std	Min	Max
fixed acidity	7.22	1.30	3.80	15.90
volatile acidity	0.34	0.16	0.08	1.58
citric acid	0.32	0.15	0.00	1.66
residual sugar	5.44	4.76	0.60	65.80
chlorides	0.06	0.04	0.01	0.61
free sulfur dioxide	30.53	17.75	1.00	289.00
total sulfur dioxide	115.74	56.52	6.00	440.00
density	0.99	0.00	0.98	1.04
pH	3.22	0.16	2.72	4.01
sulphates	0.53	0.15	0.22	2.00
alcohol	10.49	1.19	8.00	14.90

Table 1. The attribute statistics before normalisation and removal of outliers.

The goal of this project will be to find a predictor with the best regression performance.

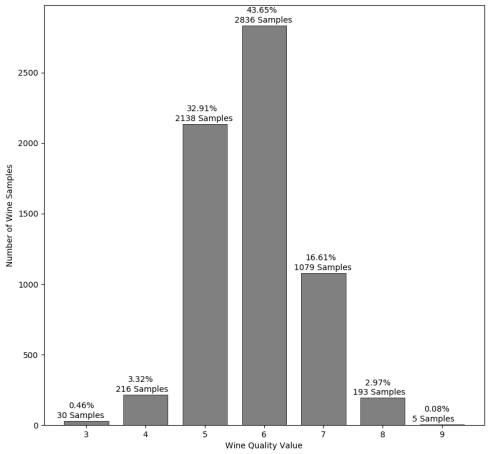

Figure 1. The histogram for both white wine and red wine qualities.

Attributes	Mean	Std	Min	Max
fixed acidity	-0.012	0.965	-2.635	4.848
volatile acidity	-0.005	0.984	-1.577	4.800
citric acid	-0.002	0.990	-2.193	4.689
residual sugar	-0.004	0.983	-1.018	4.331
chlorides	-0.049	0.740	-1.343	4.966
free sulfur dioxide	-0.008	0.968	-1.664	4.957
total sulfur dioxide	-0.001	0.998	-1.942	4.437
density	-0.004	0.979	-2.530	2.999
pH	0.000	1.000	-3.101	4.923
sulphates	-0.017	0.936	-2.092	4.898
alcohol	0.000	1.000	-2.089	3.696

Table 2. The attribute statistics after normalisation and removal of outliers.

2.1. Learning approach

For consistency all the different regressions performances of each model with varying parameters and training

cost functions will be measured using the same error metric, the Huber Loss function. This is a good loss function for this regression problem as it is robust to outliers, for if the difference between the real and predicted value is small it will be squared, if large, the absolute value will be taken.

$$L_H = \begin{cases} \sum_{i=1}^{n} \frac{1}{2}(y_i - f(x_i))^2, & for|y_i - f(x_i)| \leq \delta. \\ \sum_{i=1}^{n} \delta|y_i - f(x_i)| - \frac{1}{2}\delta^2, & otherwise. \end{cases} \tag{1}$$

This shall be used on the test error of each of the models for comparison as each model will required a different training error.

To validate which model is best they will also be tested using the Mean Absolute Deviation error metric (MAD) which is often used to test regression performance.

3. Baseline Predictors

Several baseline predictors have been implemented and trained on the data.

A linear regression program was written in Python using Tensorflow. This linear model used an iterative method to alter the weights, with multiple loss functions tested. This proved useful as a framework for the more advanced algorithms implemented later on.

For linear regression 2 basic loss functions were implemented first.

3.1. L1 loss function

The L1 loss function, least absolute deviations, tries to minimise the absolute difference between the predicted and real values. The sum of all the differences for all the samples can be described as follows:

$$L_1 = \sum_{i=1}^{n} |y_i - f(x_i)| \tag{2}$$

This is a fairly robust loss function which is not that affected by outliers.

3.2. L2 loss function

The L2 loss function, least square error, minimises the square difference between the predicted and real values. This value is much large than that of L1 loss and therefore is more affected by outliers, however, will optimise the predictor faster.

$$L_1 = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{3}$$

Notice how the predictor using the L2 loss function in 3. minimises the error much faster than the L1 loss in 2 and as outliers have been removed in data preparation the L2 loss is quite robust.

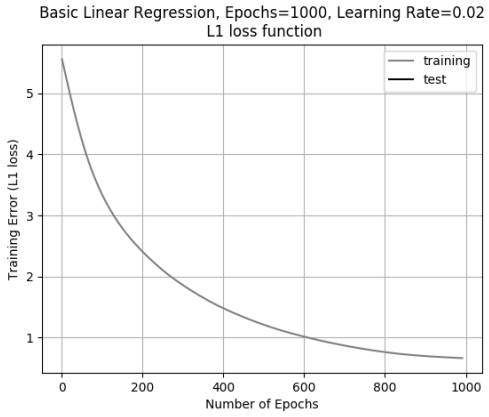

Figure 2. Training Error = $\sum_{i=1}^{n} |y_i - f(x_i)|$ (L1 loss), Epochs = 1000, Learning Rate = 0.02

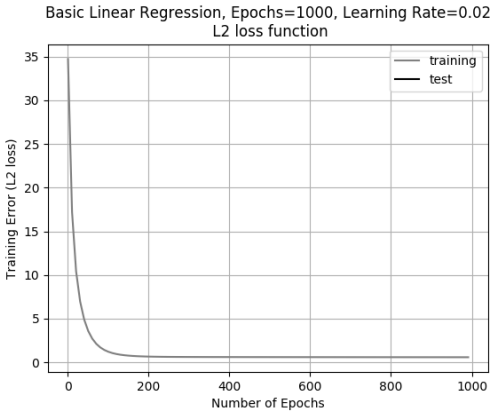

Figure 3. Training Error = $\sum_{i=1}^{n} (y_i - f(x_i))^2$ (L2 loss), Epochs = 1000, Learning Rate = 0.02

As the epochs are increased the optimiser is able to reduce the loss until no improvements can be made. The learning rate also greatly affects the rate at which the loss is minimised. However, if the learning rate is too high the predictor will not improve on the training data, as can be seen in 4.

3.3. Improvements

These relatively simple models can be improved upon by adding regularisation. Regularisation is a technique used to prevent over fitting the data. Both L1 and L2 regularisation have been implemented, being the sum of the weights and sum of the square weights respectively. The regularisation term is added to the loss function, the terms for L1 and L2 look as follows:

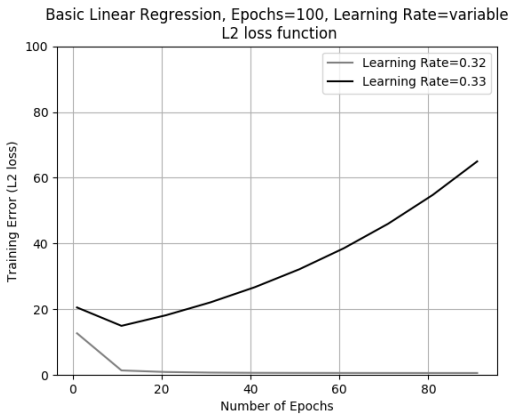

Figure 4. Training Error = $\sum_{i=1}^{n}(y_i - f(x_i))^2$ (L2 loss), Epochs = 100, Learning Rate = variable

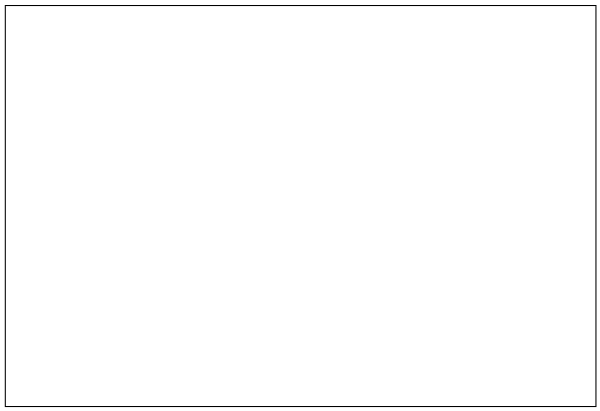

Figure 5. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

$$L_{1reg} = \lambda \sum_{i=1}^{n} |w_i| \quad (4) \qquad L_{2reg} = \lambda \sum_{i=1}^{n} w_i^2 \quad (5)$$

4. More Advanced Algorithms

4.1. Neural Network

The neural network implemented as a predictor for this project had 3 layers, 1 input, 1 hidden and 1 output. A single hidden layer was chosen as an increase in hidden layers did not result in a significant increase in predictor performance although it did lengthen training time.

The number of nodes in the hidden was chosen using a heuristic of the number being the mean of the input and output layers, in this case resulting in 6 nodes.

Each layer could be assigned when the network is constructed to have a specific activation function such as: TanH, Sigmoid, ReLU, SeLU and Softmax. It was found that during training a hidden ReLU resulted in the best out of sample test error and therefore was chosen.

4.2. Support Vector Machine

4.3. Elastic Net Regularisation

4.4. The ruler

The LaTeX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LaTeX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the command in the document preamble.) Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (e.g. this line is 095.5), although in most cases one would expect that the approximate location will be adequate.

4.5. Mathematics

Please number all of your sections and displayed equations. It is important for readers to be able to refer to any particular equation. Just because you didn't refer to it in the text doesn't mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics: `http://www.pamitc.org/documents/mermin.pdf`.

4.6. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

See The TeXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: " Frobnication has been trendy lately. It was introduced by Alpher [**?**], and subsequently developed by Alpher and

Fotheringham-Smythe [**?**], and Alpher *et al*. [**?**]."

This is incorrect: "... subsequently developed by Alpher *et al*. [**?**] ..." because reference [**?**] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al*.

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [**?**, **?**, **?**] to [**?**, **?**, **?**].

5. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

5.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high. Page numbers should be in footer with page numbers, centered and .75 inches from the bottom of the page and make it start at the correct page number rather than the 4321 in the example. To do this fine the line (around line 23)

```
%\ifcvprfinal\pagestyle{empty}\fi
\setcounter{page}{4321}
```

where the number 4321 is your assigned starting page.

Make sure the first page is numbered by commenting out the first page being empty on line 46

```
%\thispagestyle{empty}
```

5.2. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figures 5 and 6. Short captions should be centred.
Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

5.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

5.4. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [**?**]. Where appropriate, include the name(s) of editors of referenced books.

5.5. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

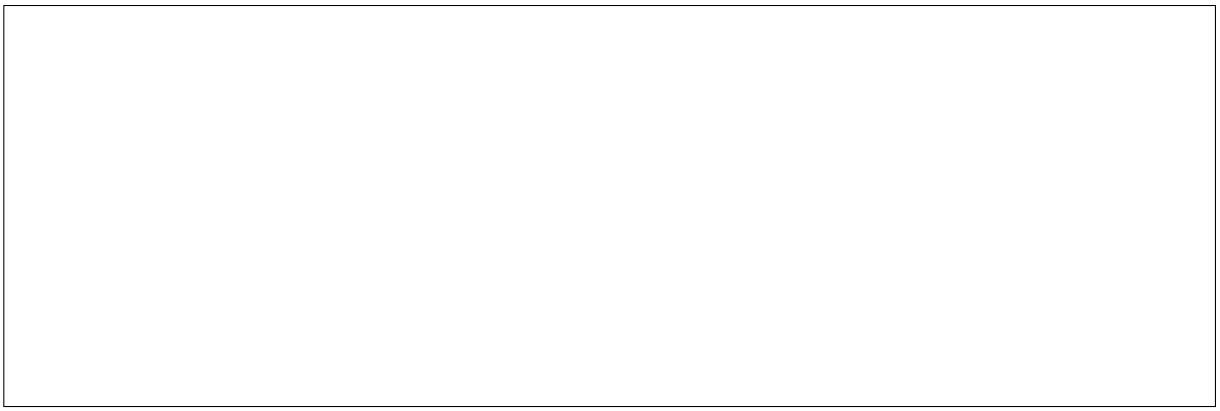

Figure 6. Example of a short caption, which should be centered.

not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

References

[1] F. A. T. M. P. Cortez, A. Cerdeira and J. Reis. Modelling wine preferences by data mining from physicochemical properties., 2019. Face and Gesture submission ID 324. Supplied as additional material `fg324.pdf`.