Final Year Project Report

Full Unit - Interim Report

Comparison of Ridge Regression Algorithm against Others When Solving Boston Housing Problem

Fabio Eugenio dos Santos de Sampaio Doria

A report submitted in part fulfilment of the degree of

BSc (Hons) in Computer Science

Supervisor: Nicolo Colombo

Department of Computer Science
Royal Holloway, University of London

Fabio Eugenio
dos Santos de Sampaio Doria

December 6, 2023

Declaration

This report has been prepared on the basis of my own work. Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count:

Student Name: Fabio Eugenio dos Santos de Sampaio Doria

Date of Submission:

Signature:

Table of Contents

Abstract

For most people buying a house will be one of their most important and expensive economic decisions that they will take in their lives.[1] Because of this it would be logical to say that being able to accurately predict the prices of said houses would be of extreme value to people. One possible way to make these predictions would be to create a machine learning model that, given a certain amount of features from each house, would be able to create an accurate prediction of their price.[2] There are a wide range of different machine learning algorithms that could be used to solve this problem. However, this project will focus on two: Ridge Regression and Decision Trees.

In this project I plan on implementing both of these algorithms and comparing their performance on one another to see which one is more effective at resolving the Boston housing problem. To measure their accuracy I will use two metrics specifically used for regression problems, mean absolute error (MAE) and root mean squared error (RMSE). MAE returns the average residual of the predictions that each model makes, this is useful as it can be directly compared against the other model for differences.[3] However, RMSE returns the square root of the average residual of the predictions. This highlights larger errors caused by the model which is practical as it helps to differentiate the performance of both algorithms.[4]

The dataset that will be used to judge the different algorithms is called the Boston housing problem datasets. The dataset is comprised of 506 entries, each having a total of 14 features which describe a property inside of the Boston Massachusetts area.[5] Two variations of this dataset will be used, one with full 14 features and another with a lower number 5. Two models will be created, one with each algorithm and each of these models will be trained and tested on both of these datasets. Their performance will then be analysed, compared, and finally conclusions will be drawn from the results in order to define the effectiveness both algorithms.

Project Specification

Your project specification goes here.

Chapter 1: **Introduction**

Since humans began roaming the earth there have always been three main things that we have needed to survive: food, water, and shelter. In modern developed civilisations food and water has become relatively easy to come by with affordable versions of both being available to most people. However, shelter seems to only become more expensive with time while income stays the same. This can be seen in the UK where the median price of residential houses increased by 14% while income decreased by 1% from 2020 to 2021, making the issue of housing one of the most prevalent ones in recent times.[6] Also, With the advent of the internet and its capability of delivering large scale of information to users, the options of houses available are in-numerous. Because of this it can be an extremely overwhelming task to look for a suitable place to live that both fits the budget but is also reasonably priced considering the aspects and features of the house. House prices also play a significant role in the economy, one of these is being linked to consumer spending. If a homeowner knows that the value of his home increases then he feels confident and is likely to spend more in goods or services or to pay off their debts. If they know the value has decreased then the opposite occurs, homeowners become less confident meaning they are likely to spend less and save more.[7] Taking all of these factors into consideration it can be said that having an accurate method of estimating house prices is of importance from an individual looking for a home all the way to governments trying to predict what economic measures they should take next.

The first step in order to generate accurate estimates is to gather and analyse data on past houses which have been sold and try to find a pattern regarding their value. Specifically between the price they were sold for and features that describe the houses, such as the size of the house, area it is located in, and so forth. This analysis used to be done through traditional statistical methods, however, recently a new approach has emerged called machine learning. Here a machine is responsible for processing and learning from the data it is given in order to make predictions on new data autonomously.[8] When it comes to predicting house prices the machine learning model is given data consisting of past sales of houses where each house sold is also accompanied with a number of features that describe it as mentioned before. The system can then learn how these features influenced the price of the houses and use this information to make predictions on new houses based only on its corresponding features.[9]

Chapter 2: **Theoretical Background**

Within machine learning there are two main methodologies utilised when trying to solve a problem, supervised and unsupervised learning. Unsupervised learning is used when the data being analysed does not need to be labelled, but instead needs to be sorted into groups by their features. On the other hand, supervised learning is concerned with providing labels to unlabelled data in an dataset, such as a list of houses without a price attached to them.[10] Within supervised learning there is once again two different types of problems that occur: classification and regression problems. Classification problems occur when the list of possible classifications is finite such as identifying a handwritten digit as its correct number. However, in the Boston housing problem the list of possible labels is infinite as they could be any price, i.e., any real number.[11] This is called a regression problem and will be the focus of the research paper.

2.1 Simple & Multiple Linear Regression

Regression as a type of machine learning problem comes from the statistical method of regression, which has the goal of determining the relationship between a dependent and independent variable(s).[12] This is then expanded to create regression models which can be formally defined as using an independent variable 'x' to predict a dependent variable 'y'.[13] With the Boston housing problem the independent variable x are the features of each house such as, per capita crime rates per town, while the dependent variable y are the prices of the houses.[14]

2.1.1 Simple Linear Regression

There are many variations of the regression algorithm that can be used in a machine learning model, however simple linear regression is arguably the most simplistic ones and serves as a base for the others. The equation used in a simple linear regression model has the form:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where y and x are the independent and dependent variable respectively. β_0 is the intercept of the line created by the equation with the y-axis, this is known as the constant term as it does not change. β_1 is the regression coefficient which defines the slope of the line. And lastly ϵ is the random error term or the random deviation which represents the difference between the predicted value and the real-life value.[15, 16]

The goal of this machine learning method is to fit the best possible line to the given data, which is achieved by finding the optimal weights of the regression coefficient β_1 and β_0, i.e., the parameters of the model. For this a cost function is used which calculates the difference between the predicted value and the expected values of every sample in the training dataset and returns it as a single real number.[17] This cost is then minimised by modifying the parameters, when the cost function cannot be further minimised then the most optimal coefficients have been achieved.[18] The most commonly used cost function for linear regression is the Sum of Squares Error (SSE) function

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

where y_i is the observed true label and $\hat{y}_i$ is the predicted label. The predicted label variable can also be substituted with the linear regression equation to get another version of the cost function.

$$SSE = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

2.1.2 Multiple Linear Regression

With the basic theoretical framework on regression being set out, it is now necessary to expand it in order to fit a practical use as a machine learning model that can be applied to real-world datasets. In reality, when creating a model their is most of the time going to be more than simply one feature that will define the outcome of the model. For this, multiple linear regression is used which has the same form as the previously seen linear regression but is extended by adding in more terms to account for the multiple features.[19]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$$

In this new equation there are now p dependent variables, each with their own β coefficient. This means that

It is also important to note that features can easily be added or removed from this model, this will be used when comparing the performance of the two algorithms on datasets with different numbers of features.

2.2 Ridge Regression

The Ridge Regression method is a different way of estimating the parameters for a multiple linear regression model, specifically in order to deal with issues that can arise from using data that suffers from multicollinearity.[20] Before looking at how Ridge Regression addresses this issue it is first necessary to understand what multicollinearity and the effects that it can have on a linear regression model.

2.2.1 Issues Caused by Multicollinearity

As said before, Linear Regression models utilise the Sum of Squares Error method to estimate the value of the regression coefficients. However, the reliability of these coefficient estimations could be severely impacted if the data being used to train the model suffers from multicollinearity, when two or more independent variables in the model are correlated.[21] This impacts the reliability of the estimations as linear regression models define the regression coefficient to represent a mean change in the dependant variable for each unit change in the independent variable when all other independent variables remain the same, and by definition this cannot occur if two or more independent variables are correlated.[22] This multicollinearity causes the model to become unstable because a small change in one variable will suddenly effect another causing a big change in the prediction. It also makes the model harder to interpret as the regression coefficients do not necessarily reflect the significance of only that specific feature.[23] Finally, it can lead to overfitting which is when a model displays a high level of accuracy on the training dataset, but performs badly when it comes to the testing set. This happens because the model is trained to specifically to the individual data points in the training set instead of the general trend of the points.Then when it tries to predict a data sample not from that set it accuracy is poor.[24]

2.2.2 Ridge Regression Solution

The Ridge Regression algorithm solves these issues caused by multicollinearity by applying a processes to the model called regularisation, specifically L2 Regularisation. Regularisation is the process of 'regularising' the regression coefficients by trying to make them as small as they need to be.[25] The effect this has on the model is of generalisation, meaning that it wont be as accurate with its predictions in the training set, but it will work better in the long run with predictions on new unseen data. In more precise terms it can be said that a small amount of bias is added into the model which in return decreases the amount of variance which causing the poor performance.[26]

L2 regularisation achieves this by enhancing the SSE equation by adding a penalty term to the end of it. This term is the summation of squared weights of each feature, multiplied by 'a' that defines how 'harsh' the penalty term should be. Finding the optimum value for 'a' is critical as it defines the overall performance of the equation, if it is zero then the model once again uses just SSE and the higher 'a' is the more generalised and less accurate the model will be.[27]

$$SSE + a||\beta||^2 = \sum_{i=1}^{n}(y_i - \beta x_i)^2 \ + a\sum_{j=1}^{P}\beta_j^2.$$

This equation is then minimised and after this is complete the most optimal coefficients will be found that will help prevent overfitting in the model.

2.3 Random Forest

Chapter 3: **Methodology**

Chapter 4: **Headings**

Your report will be structured as a collection of numbered sections at different levels of detail. For example, the heading to this section is a first-level heading and has been defined with a particular set of font and spacing characteristics. At the start of a new section, you need to select the appropriate LaTeX command, `\chapter` in this case.

4.1 Second Level Headings

Second level headings, like this one, are created by using the command `\section`.

4.1.1 Third Level Headings

The heading for this subsection is a third level heading, which is obtained by using command `\subsection`. In general, it is unlikely that fourth of fifth level headings will be required in your final report. Indeed it is more likely that if you do find yourself needing them, then your document structure is probably not ideal. So, try to stick to these three levels.

4.2 A Word on Numbering

You will notice that the main section headings in this document are all numbered in a hierarchical fashion. You don't have to worry about the numbering. It is all automatic as it has been built into the heading styles. Each time you create a new heading by selecting the appropriate style, the correct number will be assigned.

Chapter 5: **Presentation Issues**

5.1 Figures, Charts and Tables

Most final reports will contain a mixture of figures and charts along with the main body of text. The figure caption should appear directly after the figure as seen in Figure 5.1 whereas a table caption should appear directly above the table. Figures, charts and tables should always be centered horizontally.

Figure 5.1: Logo of RHUL.

5.2 Source Code

If you wish to print a short excerpt of your source code, ensure that you are using a fixed-width sans-serif font such as the Courier font. By using the `verbatim` environment your code will be properly indented and will appear as follows:

```
static public void main(String[] args) {
  try  {
    UIManager.setLookAndFeel(UIManager.getSystemLookAndFeelClassName());
  }
  catch(Exception e) {
    e.printStackTrace();
  }
  new WelcomeApp();
}
```

Chapter 6: **References**

Use one consistent system for citing works in the body of your report. Several such systems are in common use in textbooks and in conference and journal papers. Ensure that any works you cite are listed in the references section, and vice versa.

Chapter 7: **Project Information and Rules**

The details about how your project will be assessed, as well as the rules you must follow for this final project report, are detailed in the project booklet.

You must read that document and strictly follow it.

Bibliography

[1] L. R. Weinstock, "Introduction to u.s. economy: Housing market," *Congressional Research Service*, Jan. 2023.

[2] P. Herman, "The importance of price prediction," *Future Processing*, Mar. 2023.

[3] P. Schneider and F. Xhafa, "Chapter 3 - anomaly detection: Concepts and methods," in *Anomaly Detection and Complex Event Processing over IoT Data Streams* (P. Schneider and F. Xhafa, eds.), pp. 49–66, Academic Press, 2022.

[4] S. Olumide, "Root mean square error (rmse): What you need to know," *Arize*, Aug 2023.

[5] V. Roman, "Root mean square error (rmse): What you need to know," *Towards Data Science*, Jan. 2019.

[6] C. Smith, "Housing affordability in england and wales: 2021," *Census 2021*, Mar. 2022.

[7] "Housing affordability in england and wales: 2021," *Bank of England*, Mar. 2020.

[8] M. Chatterjeeh, "Data science vs machine learning and artificial intelligence: The difference explained (2024)," *Great Learning*, Nov. 2023.

[9] I. Ake, "Combining machine learning models to predict house prices," *Solent University*, Sep. 2022.

[10] V. Kanade, "What is machine learning? definition, types, applications, and trends for 2022," *Spice Works*, Aug. 2022.

[11] V. Vovk, "Chapter 2: Introduction to machine learning and nearest neighbours." `https://moodle.royalholloway.ac.uk/pluginfile.php/188746/mod_resource/content/27/02_1.pdf`, Sep. 2023.

[12] B. Beers, "What is regression? definition, calculation, and example," *investopedia*, Mar. 2023.

[13] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," *Journal of Applied Science and Technology Trends*, vol. 1, pp. 140–147, Dec. 2020.

[14] S. Gupta, "Boston house price prediction based using support vector regressor," *enjoy algorithms*.

[15] S. Rong and Z. Bao-Wen, "The research of regression model in machine learning field," in *MATEC Web of Conferences*, vol. 176, p. 01033, EDP Sciences, 2018.

[16] S. Glen, "Error term: Definition and examples," *Statistics How To*, Nov. 2020.

[17] K. Krzyk, "Cost function of linear regression: Deep learning for beginners," *Built In*, Jul. 2022.

[18] S.-J. Kim, S.-J. Bae, and M.-W. Jang, "Linear regression machine learning algorithms for estimating reference evapotranspiration using limited climate data," *Sustainability*, vol. 14, no. 18, p. 11674, 2022.

[19] M. Tranmer and M. Elliot, "Multiple linear regression," *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, vol. 5, no. 5, pp. 10–11, 2008.

[20] G. C. McDonald, "Ridge regression," *WIREs Computational Statistics*, vol. 1, no. 1, pp. 93–100, 2009.

[21] A. Alin, "Multicollinearity," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 3, pp. 370–374, 2010.

[22] J. Frost, "Multicollinearity in regression analysis: Problems, detection, and solutions," *Statistics By Jim*, 2017.

[23] S. Wu, "Multicollinearity in regression," *towardsdatascience*, May. 2020.

[24] J. Nagidi, "How to handle overfitting in deep learning models," *Dataaspirant*, Aug. 2020.

[25] P. Gupta, "Regularization in machine learning," *Towards Data Science*, Nov. 2017.

[26] C. Maklin, "Machine learning algorithms part 11: Ridge regression, lasso regression and elastic-net regression," *Medium*, Dec. 2018.

[27] K. Kargin, "Ridge regression fundamentals and modeling in python," *Medium*, Apr. 2021.