Ridge Regression — A graphical tale of two concepts

Tanveer Hurra · Follow
Published in Towards Data Science · 7 min read · Apr 17, 2020

12

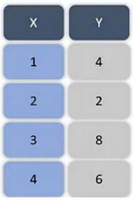
Photo by Abraham Osorio on Unsplash

Regression is most probably the first machine learning algorithm that one learns. It is basic, simple and simultaneously a very useful tool that solves a lot of machine learning problems. This article is about Ridge Regression, a

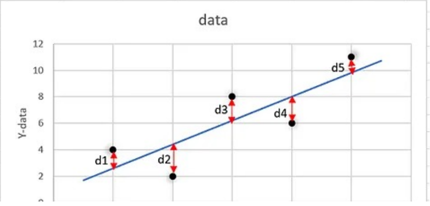
To understand this concept, read all the four parts very carefully. Now, let's start:

1. **Linear Regression**

Suppose we are given a two-dimensional data set, shown and plotted below, and we intend to fit a linear model into it. The black dots in the graph show the data points that are shown in the left table. The blue line is the linear model fitted into the data & the red arrows show the difference in the predictions & actual dependent variable values (Y).



X	Y
1	4
2	2
3	8
4	6



In a linear regression model, a line is fitted into the data in such a way that d1^2 + d2^2 + d3^2 + d4^2 + d5^2 is minimized i.e. the square of the residuals (the difference between the actual values and predicted values) is minimized. In more general form it can be represented as shown below:

$$\sum_{i=1}^{n}(yi - Bo - BiXi)^2$$

The quantity is called the residual sum of squares
(RSS), here yi represents the predicted value of the dependent variable. The method of finding the linear model in this way is called the ordinary least squares method. The fitted model is of the form as given below:

$$Y = Bo + B1X$$

Bo in the above equation is the intercept and **B1** is the slope of the model, **X** and **Y** is the independent and dependent variable respectively.

2. Feature Selection

In any machine learning problem, we are given predictors also called features or independent variables and based upon the data provided we need to understand the relationship of these variables to the response variable also called the dependent variable. In simple terms we need to find out the below relationship:

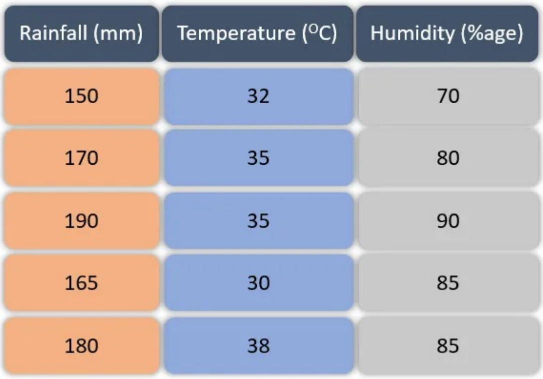
$$y = f(x1, x2, x3, \ldots\ xp)$$

We have **y** as the response variable and **p** predictors (x1, x2, … xp) and the above equation represents the actual functional relationship between the two. In machine learning we need to find this relationship as accurately as possible. The search for the model brings a number of challenges with it like insufficient or missing data, missing predictors, irrelevant predictors, correlation among the predictors, wrong format of data etc. All the challenges that we face in machine learning are mitigated through different methods. The challenge that we will address in this topic is ”irrelevant predictors”. I will illustrate this problem with an example.

Consider understanding the relationship between the amount of rainfall with temperature, humidity, geographical location and hair colour of the people in the area. The intuition suggests that the rainfall occurrence can have a relationship with the first three predictors but there seems to be no logical link between the rainfall and hair colour of the people. Given the data set, if we try to fit a regression model into it, the model will adjust itself with the hair colour data too, which is wrong. Ideally, our model training method should eliminate unnecessary predictors or give them less weightage at least. Similarly, if there is a correlation among predictors (as humidity has with geographical location, coastal areas are more humid etc.), the model should learn that too. The process of eliminating or reducing weightage of unnecessary predictors is called feature selection. This concept is important in the current scenario as ridge regression directly deals with it.

3. Parameter calculation in regression

Although the process of least squares method in regression has already been discussed, it’s time to understand the concept graphically. Consider the data set as given below. The system of rainfall amount with temperature and humidity is represented by the data:



Rainfall (mm)	Temperature (^{O}C)	Humidity (%age)
150	32	70
170	35	80
190	35	90
165	30	85
180	38	85

If we wish to fit a model in the above data through regression, we need to fit a linear equation of type **Y = Bo + B1 X1 + B2 X2** in it or more simply we need to calculate **Bo**, **B1** and **B2**. Here **X1** will be temperature and **X2** will be humidity and **Y** would be the amount of rainfall. Let's forget **Bo** (the intercept) for a while and choose arbitrary values for **B1** and **B2**, say **B1** = 1 & **B2** = 2. These are not the desired values but randomly chosen to understand a concept. Also, let's take the value of zero for **Bo**. So now, we have parameter values with us, let's predict the rainfall values and calculate RSS (sum of square of errors). The below table shows the required calculations:

As evident by the above table, for **B1** = 1 and **B2** =2, the RSS value is 3743. Before concluding anything here, let's assume a different value set for regression parameters. Let B1 = 3.407 and B2 = 1 and let's keep ignoring **Bo**. If we make prediction calculations again & calculate RSS too, you will find out that it would again come out to be 3743 (almost). The point I want to make here is that there are different values of regression parameters for which the residual sum of squares is constant. These points if plotted will generate a plot as shown below:

The above plot is for a system of two predictors, more number of predictors will increase the dimension of the above plot accordingly. We initially assumed **Bo** to be zero, but for any value of **Bo**, the plots will still be the same, only shifted upwards or downwards depending upon the sign & magnitude of **Bo**.

The above plots are called **cost contour plots of regression**. Each contour or loop is plotted between parameters **B1** and **B2** and represent a constant RSS value. In regression, we aim to find out the value represented by the dot at the centre, which is both unique and represent minimum RSS.

4. Ridge Regression

Ridge regression is a modification over least squares regression to make it more suitable for feature selection. In ridge regression, we not only try to minimize the sum of square of residuals but another term equal to the sum of square of regression parameters multiplied by a tuning parameter. In other words, in ridge regression we try to minimize the below quantity:

The first term in the above expression is the sum of squares of residuals and the second term is what is specially added in ridge regression. Since this is a special term introduced in ridge regression let's try to understand it further. For a data set with two predictors, it will be *a* (B1^ 2+ B2^ 2), where *a* is the tuning parameter. It is also called the penalty term as it puts a constraint on the least-squares method of regression. In the quest of minimizing it, it is constrained to a particular value, depicted by the below equation:

Look at the above equation carefully, it is the equation of a shaded circle, with square of radius equal to ***s/a*** *(the constraint)*, a part of which is shown below:

Combining the above graph with the cost contour graph will result in the graph as shown below:

The above graph gives the idea of ridge regression. It is where the least-squares condition meets the parameter constraint or penalty condition. The radius of the circle representing constraint directly depends upon the tuning parameter (a). Latger the value of tuning parameter, smaller the circle, higher the penalty. You can imagine it directly with the help of the above graph. Larger the value of tuning parameter, smaller will be the circle, closer to the origin will be the meeting point of two graphs, hence smaller the values of regression parameters.

Conclusion:

In ridge regression, finding the parameters corresponding to the minimum residual sum of squares is not what is sought. A constraint is put on the parameters to put a check on them and hence not allowing them to grow. This condition makes sure that different parameters are given weightage differently and hence becomes an important tool for feature selection. Please note that in ridge regression no parameter of any predictor is made zero but parameter weightage is varied

That is all for ridge regression. Do post your comments/suggestions. For any query regarding the topic, you can reach me on **LinkedIn**.

Further Read:

- **Lasso Regression**
- **Subset Selection**

Thanks,

Have a nice time 😊

· · ·

Originally published at https://www.wildregressor.com on April 17, 2020.

Machine Learning Data Science Technology Artificial Intelligence

12

Written by Tanveer Hurra

Follow

244 Followers · Writer for Towards Data Science

MBA — Business Analytics|| https://in.linkedin.com/in/tanvirhurra || https://www.teastatistic.com/

More from Tanveer Hurra and Towards Data Science

Tanveer Hurra in Towards Data Science

DBSCAN — Make density-based clusters by hand

A dumb approach to DBSCAN, an unsupervised machine learning algorithm

7 min read · May 4, 2020

67 2

Cristian Leo in Towards Data Science

The Math Behind Neural Networks

Dive into Neural Networks, the backbone of modern AI, understand its mathematics,...

✦ · 28 min read · Mar 29, 2024

3K 20

Tim Sumner in Towards Data Science

A New Coefficient of Correlation

What if you were told there exists a new way to measure the relationship between two...

10 min read · Mar 31, 2024

2.8K 35

Tanveer Hurra in The Startup

Linear Discriminant Analysis — Basics with hands-on practice

Basics explained with hands-on practice over an example to illustrate it further

6 min read · Mar 26, 2020

62

See all from Tanveer Hurra

See all from Towards Data Science

Recommended from Medium

Tim Sumner in Towards Data Science

A New Coefficient of Correlation

What if you were told there exists a new way to measure the relationship between two...

10 min read · Mar 31, 2024

2.8K 35

Jyoti Dabass, Ph.D in Python in Plain English

Friendly Introduction to Deep Learning Architectures (CNN, RN...

This blog aims to provide a friendly introduction to deep learning architectures...

8 min read · Apr 2, 2024

680 6

Lists

Predictive Modeling w/ Python
20 stories · 1113 saves

AI Regulation
6 stories · 417 saves

ChatGPT prompts
47 stories · 1451 saves

Natural Language Processing
1386 stories · 882 saves

Rosaria Silipo in Low Code for Data Science

Is Data Science dead?

In the last six months I have heard this question thousands of time: "Is data science…

6 min read · Mar 11, 2024

1.7K 36

Learnbay_Official

Pytorch vs. Tensorflow: Major Difference Among Deep Learning

Comprehend Significant Differences Between Pytorch and Tensorflow for Deep Learning

12 min read · Nov 21, 2023

78

Sertis

Deep Learning Roadmap: Step-by-Step from Zero to Hero

Whether you are a data analyst, data engineer, software developer, or simply…

4 min read · Nov 14, 2023

103

Marco Del Pra

Generative Adversarial Networks

Generative Adversarial Networks (GANs) have garnered significant attention in the…

14 min read · Oct 30, 2023

22

See more recommendations