BYTE PANACHE COMPETITION

FOUNDATIONS OF MODERN MACHINE LEARNING

OBJECTIVES

The objective of this competition was to use the concepts of **regression and regularization** we have learnt to predict early Covid-19 cases. WE used **linear regression, polynomial regression and ridge regression** to obtain a reasonably good estimate of the future cases.

MATERIALS & METHODS

The following methods were used to train the models:

- Linear Regression
- Polynomial Regression
- Ridge Regression
- Ridge Polynomial Regression

The following equations were used for performance analysis:

- Mean Square Error

$$\sum_{i=1}^{m}(x_i - \vec{x})^2/m \quad (1)$$

- Mean Absolute Error

$$\sum_{i=1}^{m}|x_i - \vec{x}|/m \quad (2)$$

- R2 Score

$$1 - ((\sum_{i=1}^{m}(x_i - \vec{x})^2/m)/(\sum_{i=1}^{m}|x_i - \vec{x}|)/m) \quad (3)$$

- m = total number of examples
- xbar = mean of all examples of x

INTRODUCTION

We are given a dataset pertaining data related to the COVID-19 cases occurred between the dates *22/01/2020* and *31/03/2020*. The data included the number of confirmed and recovered cases, deaths,area of occurrence of the cases under discussion and also their date of observation.

RESULTS 2

The values of the performance metrics obtained after training and building the model on the whole dataset are tabulated below

Model	MSE	MAE
Linear Reg	164288029427	368049.3
Polynomial Reg	1314905592	31989.7
Ridge Reg	164440629513	368251.9
Ridge Poly Reg	352565817	15796.1

Table 1: For Whole Dataset

The values of the performance metrics obtained after training and building the model on the dataset values consisiting of Asian Countries are tabulated below.

Model	MSE	MAE
Linear Regression	181730284.9	10324.9
Polynomial Regression	450761594.1	15424.4
Ridge Regression	182125200	10329
Ridge Poly Reg	391656065	17728

Table 2: For only Asian Countries

RESULTS 1

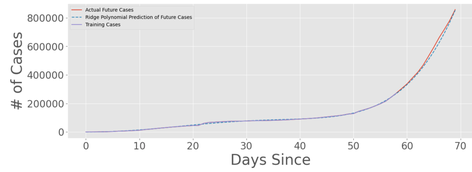


Figure 1: Ridge PolyReg for whole dataset

On the whole dataset after performing linear, polynomial, ridge, poly ridge regression, and plotting the results, we observe that polynomial ridge regression does perform the best on the provided dataset. We observe that China accounts to about 40 percent of the cases which is the highest amount by a single country. This can be confirmed by going through the pie chart in Part-3 of the colab notebook.

For the next part we considered only Asian countries to predict their number of cases. Even here after applying linear, polynomial, ridge, poly ridge regression, and plotting the results, we observe that polynomial ridge regression does outperform its counterparts on the provided dataset.

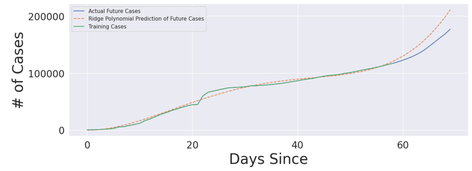
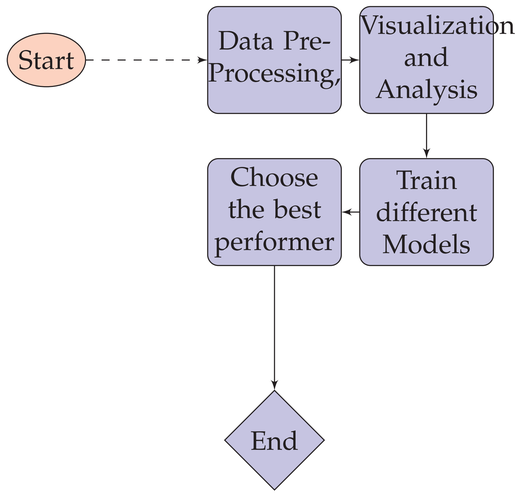


Figure 2: Ridge PolyReg for only Asian countries

CONCLUSION



- Polynomial Regressors do seem to fit better due to the non linear nature of the data.
- Due to the time series nature present in the data we should be mindful of not shuffling the data while splitting.
- Polynomial Ridge Regressor seems to be working the best of all the regressors used on this dataset due to the non-linearity in the data and also the regularization nature present in its nature thus reducing the possibility of overfitting.

REFERENCES

[1] Sebastian Raschka, Yuxi Liu, and Vahid Mirjalili. *Machine Learning with PYTORCH and Sci-Kit Learn*. Packt, 1st edition, 2022.

[2] Medium. Various Titles. *Towards Data Science*, 2010-present.

FUTURE RESEARCH

The current predictions was based on the idea that the trend followed would be only increase in the number of cases, but the reality was not that there was a pattern of increase and decrease in the total number of cases over time.

It would be more beneficial to train models with data which correctly represents these trends so as to make a more accurate prediction. Also implementing other algorithms as DT or RF Reg would also help make a more accurate predictions.

CONTACT INFORMATION

Email psvkaushik@gmail.com
Phone +91 (900) 091 5599