Prakhar Saxena

Professor Edward Kim

CS 383 - 001

May 14, 2020

Honours Project: L1 Normalisation

In machine learning, one of the critical tasks is selecting important features that are to be used in a Machine Learning model. These features are often hand-selected by the developer; however, their relative importance to the task at hand is unknown. In this project I used a mathematical regularisation technique, called L1 Normalisation, to identify the features that are most important to a machine learning model.

Overfitting

"The production of an analysis which corresponds too closely or exactly to a particular set of data and may, therefore, fail to fir additional data or predict future observations reliably."

"An overfitted model is a statistical model that contains more parameters than can be justified by the data."

Regularisation

In the simplest terms Regularisation is a process of adding information to solve overfitting in a Machine Learning Model. It artificially discourages complex equations and

solutions, even if they fit perfectly with the training/observed data. Explanations or solutions typically of those complexities do not bode well in generalising a real-world test data, because it fits too closely to the training data leaving any space of freedom for the test data.

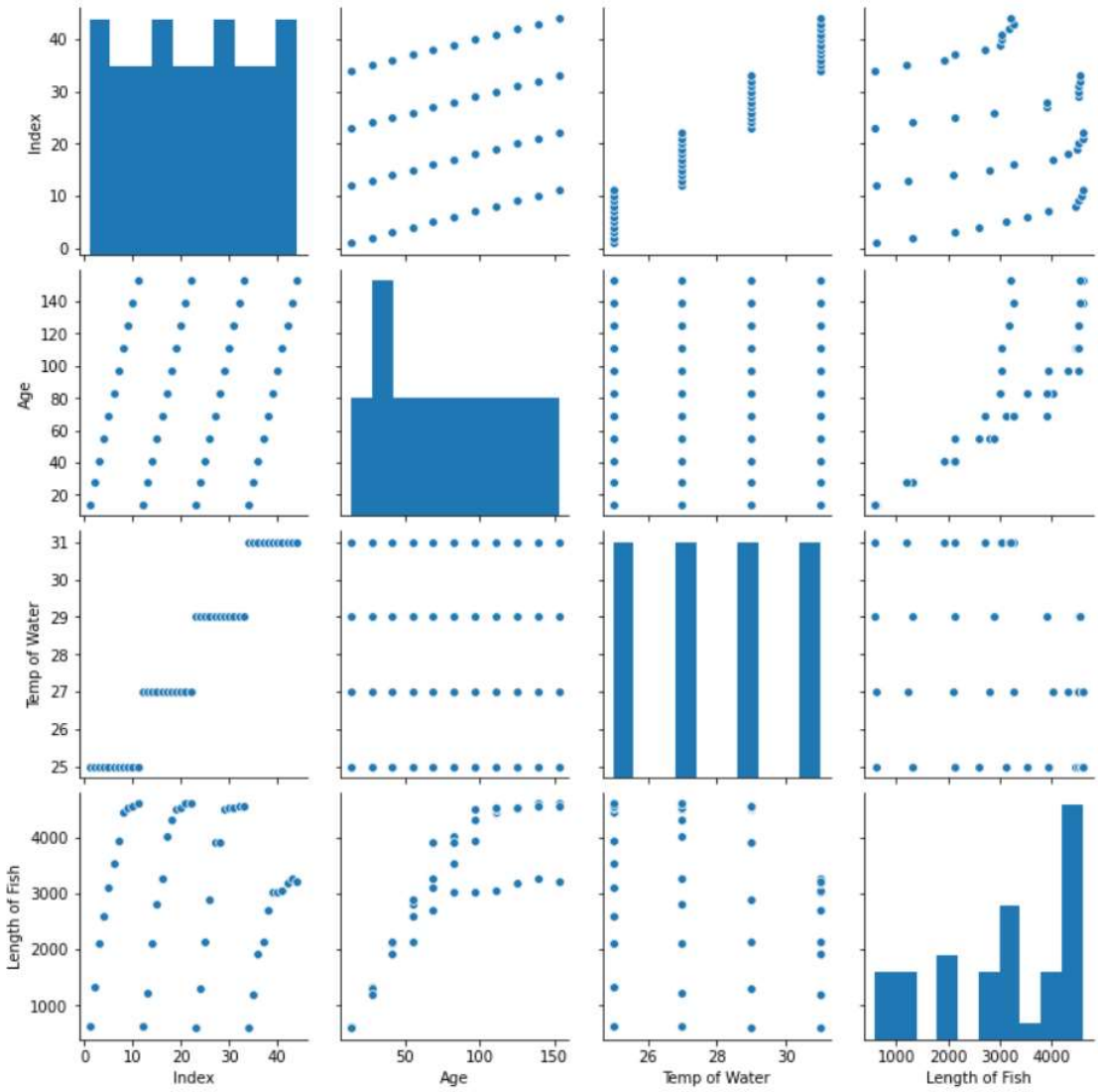
Today there are countless libraries that can enable us to regularise a dataset. I used the most popular one called Lasso in sklearn.linear_model. I also implemented my own version.

Dataset

Before I can talk about implementing any model or regularisations, I will briefly talk about what the dataset consisted of. We were given a simple comma-separated-file(.csv)

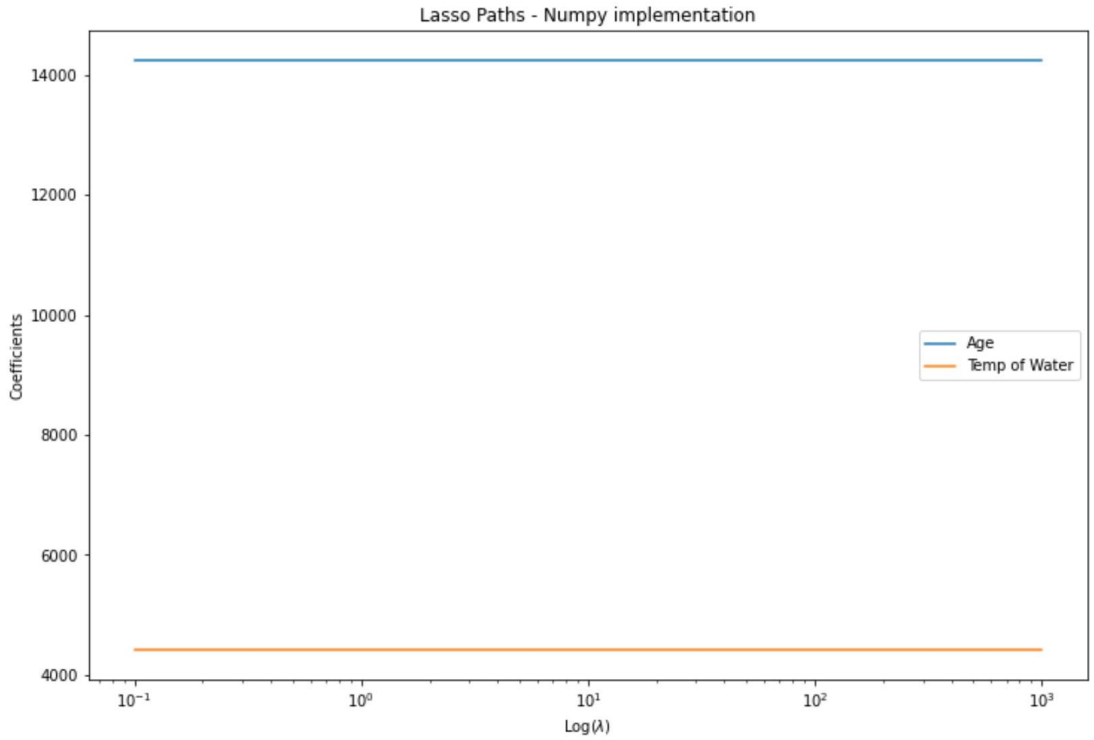
Schema: Index, Age, Temp of Water, Length of Fish

The target attribute in this dataset is the Length of the Fish, and we use other attributes to predict and/or compute it. There are only 44 entries in the dataset, and I believe that is because this was the part of the first assignment in this course.



Doing it

In the final implementation, I initialised the Lasso object with alpha, which is a hyperparameter change according to the data, as 3. I tried variations 1 and two before that. On fit I got the coefficient Estimate to be -28.19, 27.486, 0. However, that was with the open library version.



In the conclusion for this run there were a total of two features. For features in the chart above, the Age is a far more critical attribute for our models than the Temperature of Water.

References

https://www.lexico.com/definition/overfitting

Everitt B.S., Skrondal A. (2010), *Cambridge Dictionary of Statistics*

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html