Remediation of Overfit of Continuous Regression Models Across Complexity Through Ordinal Classification Training

McKade Thomas

August 30, 2022

Spring 2022

ABSTRACT

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

1. INTRODUCTION

A common topic in data science and machine learning is the ability to classify data into different categories or groups based on patterns of information. In some cases, there is a natural organization to these categories such that the labels assigned to observations have an order to them. This task, commonly referred to as ordinal classification, uses the information contained in this natural order to provide predicted labels that represent a hierarchy in the data.

Often times in this framework, the labels assigned to observations in ordinal classification provides only surface level insight for the phenomenon due to the fact that these labels are coarse. In this situation, a continuous measure can provide further insight not contained in the labels themselves.

When a set of existing binned labels exists for training an ordinal classification model, the information gained by this model can also be used to learn a related continuous measure. So long as the relationship between this continuous measure and the original binned labels is well established, a connection can be drawn by a second regression model that predicts the continuous measure based on the information learned by the original classification model.

While this approach broadens the understanding and patterns of the original binned labels, there exists the concern that the regression model, predicting a continuous measure related to the ordinal labels, will learn the set of labels and the noise associated with them quickly, leading to overfitting. Additionally, the more complex a continuous regression model becomes, the model begins to generalize more and more poorly to new data.

2. LITERATURE REVIEW

Points to discuss:

1. using binned labels for learning continuous measures
2. affect of sample size on overfit
3.

The task of ordinal classification is a subclass of classification in machine learning that has not been explored as exhaustively as other subclasses. It has been found that...

3. ORDINAL CLASSIFICATION

4. OVERFITTING TO BINNED LABELS Using data obtained from the UC Irving Machine Learning Repository from a small-scale steel industry in South Korea, this trend was explored. The data presents information on energy consumption for producing several types of coils, steel plates, and iron plates. This dataset was selected due to its large amount of observations, continuous response variable, and relatively small number of predictors.

The first step in understanding this phenomon was to create synthetic labels based on the original continuous response. Using the distribution of usage per Kwh was split into 5 categories resulting in the following splits: —insert table—

This created a set of synthetic ordinal binned labels (1-5) correlated with the original continuous response of Usage per Kwh.

5. EFFECTS OF SAMPLE SIZE

complexity_monte_carlo_10000.png

6. AVOIDING OVERFIT BY ADDING NOISE

7. CONCLUSIONS

REFERENCES