COMP 551 - Mini Project Two  
Group 23 - Dean Meluban, Javier Ordenes Alegria, Ryan Wilson

**Abstract**

In this mini project, our group implemented and investigated the performance of a multi-class logistic regression algorithm, specifically softmax regression, in tandem with a mini-batch optimization algorithm: gradient descent with momentum. This was compared to KNN, another regression classifier technique studied in this course. The datasets used were Scikit-Learn Digits Dataset and OpenML’s [wine-quality-white](https://www.openml.org/d/40498) dataset. Results showed that using 5-fold cross validation and GridSearch CV techniques, softmax regression were \_\_\_\_\_\_\_\_ in comparison to KNN regression accuracy. Different parameters of the optimization function resulted in different performances, as \_\_\_\_\_\_\_\_\_ parameters yielded higher performances, as visualized through training and validation curves.

Introduction

**Datasets**

The datasets used were the Scikit Learn Digit Dataset and the Open ML ‘wine-quality-white’ dataset. The digits dataset has 10 classes, and 1797 samples, with each feature being an integer between 0 and 16. The wine dataset had 7 classes and 4898 samples, with each of its 11 features corresponding to a different property of the wine (fixed acidity, citric acid, pH, etc.)[[1]](#footnote-1) Normalization on the digits dataset was underdone to help refine the model and optimization: specifically mean-centering and setting the Euclidean norm to 1.0). To complete the 5-fold CV technique, the total rows of each dataset were first shuffled, then split into 5 distinct entities in order for training and testing to be completed correctly. Certain variables were one-hot encoded to convert the categorical data into integer data. It is important to note that in the OpenML wine dataset, the attributes are dependent on each other, and may be correlated, hence in some circumstances it may be applicable to utilize feature selection in regression tasks.

Results

Using a shuffling and 5-fold cross validation technique of the digits

Discussion/Conclusion Ideas

Perhaps we should have done PCA or remove some features b/c they might be correlated and independent

1. https://www.openml.org/d/40498 [↑](#footnote-ref-1)