Abstract

Implement five different classifications. These models include Logistic regression, Decision trees, Support vector machines, Adaboost, and random forest, and will be applied to two text-based data sets. We have compared the accuracy and efficiency of all five models on the datasets. The results showed that trees had a lower efficiency and sometimes lower accuracy than the linear-based models. We believe this to be the result of numerous amounts of branches, due to the classification style of tree-based models.

Introduction

Using the two datasets we were able to get an idea of how different classification methods compare to each other in terms of their efficiencies and accuracies. We learned that with dataset 1 linear models tended to not only be more efficient, but also more accurate in most cases. For example, Logistic regression and SVM were both more accurate and efficient than random forest and decision trees. Although Adaboost was the most inaccurate of all even though it was a linear model. This allowed it to be more efficient than both tree models, but efficiency is worth nothing if the model lacks accuracy.

Related work (4+ sentences)

Summarize previous literature related to the multi-class classification problem and text classification.

Dataset and setup (3+ sentences)

Briefly describe the dataset and explain how you extracted features and other data pre-processing methods that are common to all your approaches.

Proposed approach (7+ sentences)

Briefly describe the different models you implemented/compared and the features you designed, providing citations as necessary. If you use or build upon an existing model based on previously published work, it is essential that you properly cite and acknowledge this previous work. Discuss algorithm selection and implementation. Include any decisions about training/validation split, regularization strategies, optimization tricks, setting hyper-parameters, etc. It is optional to provide detailed derivations for the models you use, but you should provide at least a few sentences of background (and motivation) for each model.

Results (7+ sentences, possibly with figures or tables)

Linear regression had an accuracy of 78 percent.

	Dataset 1	
Logistic Regression	0.7801380775358471	
Decision trees	0.44122700134031917	
SVMs	0.7991237387148168	
Ada boost	0.474907063197026	
Random Forest	0.6991502920870951	

Discussion and Conclusion (3+ sentences)

Summarize the key takeaways from the project and possibly directions for future investigation.

Statement of Contributions (1-3 sentences)

State the breakdown of the workload.