COMP 550: Programming Assignment 1

Jan Tiegges

September 29, 2023

1 Introduction

In this assignment, the use of different linear classifiers as well as the impact of different choices regarding preprocessing and hyperparameters was investigated based on a toy experiment. The task was to classify short statements about animals whether it is a true or an imaginary one. The primary research question was to find out whether linear classifiers are suitable for such tasks and which effect the choice of a classifier has.

2 Methodology

The experimental data was generated with ChatGPT by providing clear guidelines and examples. 230 statements were obtained for 23 animals (10 facts and fakes each) and the combination with additional data from students resulted in 1030 statements for each category. For preprocessing, the data was split into individual tokens and then optionally lemmatization and porter stemming were applied and the stop words were removed. For feature extraction, the data was split into an 80/20 train/test split and processed using a count vectorizer with unigrams. Classification was performed using Naive Bayes (NB), Logistic Regression (LR) and Linear Support Vector Machines (SVM). For all preprocessing options, a 5-fold cross-validation (cv) optimized the regularization parameter C for LR (ranged from 0.1 to 5) and SVM (ranged from 0.1 to 100) and the smoothing parameter α for NB (ranged from 0.1 to 2). The LR solver was "Saga" with a maximum number of 5000 iterations. A further 5-fold cv was then performed to find the best preprocessing options for each model.

3 Results

The performance range between the best and worst models across different preprocessing options was very minimal. Notably, the NB model with the poorest cv training score of 0.934 outperformed on the test set with a score of 0.944, compared to the NB model with the highest cv train score (0.947) but a test score of 0.937. The latter achieved the highest cv train score among all models, yet the lowest test score. For both LR and SVM models, the least effective results were observed when no preprocessing was applied. Interestingly, the NB model exhibited its highest performance with zero preprocessing. Considering the test set performance, the LR model stood out with the highest score of 0.954, closely followed by the SVM with 0.949. In terms of sensitivity to preprocessing options, SVM showed the least variability, with a small 0.005 range between its best and worst model performance. In contrast, the LR model had a more significant range of 0.012. The top-performing models for LR and SVM used lemmatization. Furthermore, SVM benefited from stemming, while the LR model achieved better results with the removal of stop words.

4 Discussion

The strong results show that linear classifiers are suitable for such use cases. LR showed the strongest performance and also the greatest sensitivity to preprocessing, which may be due to its reliance on probabilistic estimates. SVM showed robustness regardless of preprocessing, which may be due to its margin-based decision boundary that could effectively capture the differences between real and fake statements. The differences in NB performance between train and test set could be due to the assumption of feature independence, especially when the fake statements have individual features that are not present in the training dataset. Although the results are promising, one must be cautious in interpreting them, especially due to the limited dataset. Since the data is from GPT-3.5, the inherent structure and patterns may have helped the classifiers. The underlying assumption of classifying between fact and fake based on language cannot be generalized. However, in the cases where the invented facts use a completely different language (e.g. dancing, Olympics), while the facts are rather neutral scientific facts, this assumption is quite plausible. Moreover, the link between animal and fact was not considered.

5 Conclusion

We observed how strong linear classification model work on simple data. However, even if the assumptions were reasonable here, the generalization to other data is very limited. Future work should dig deeper into the nuances of the models and account for data variations.