CSC 820
Homework 10
Andrew Dahlstrom
4/22/2024

This project provides an evaluation of a Python's SKlearn logistic regression model trained to classify text samples by authorship. I experimented with a variety of hyperparameter settings across different cross-validation settings. The performance for each variation was compared using GridSearch and the optimal combination was selected to develop the model. An investigation into feature importance and misclassification patterns was also conducted. The model incorporates text features and engineered numerical attributes such as word counts and sentence lengths, number of nouns, adjectives and verbs.

Cross-Validation (CV) K-fold size and Hyperparameter tuning with GridSearch
Three cross-validation settings were employed to assess model performance: 2-fold, 10-fold, and 20-fold.

The following hyperparameters were adjusted during the cross-validation comparison:

- Max Document Frequency (max_df): Tested at 0.9 and 0.95 to exclude terms appearing in more than these thresholds (90% of documents, 95% of documents)
- TF-IDF N-gram Range: Examined the impact of using unigrams ((1,1)) versus a combination of unigrams and bigrams ((1,2)).
- Regularization Strength: Values of 0.1 and 10

The results indicated an improvement in model performance with increasing numbers of folds:

- 2-fold CV resulted in an accuracy of approximately 76.89%.
- 10-fold CV showed a better accuracy of 80.40%.
- 20-fold CV reached the highest accuracy of 80.67%.

These results suggest that higher fold settings provide a more reliable estimate of model performance, likely due to a more comprehensive utilization of the dataset for training and validation in each cycle. The improvement from 10-fold to 20-fold only achieved about a .27% improvement indicating that the increased computational complexity of more than 10-fold might not be worth the small performance improvement.

In addition, the optimal hyperparameter settings for each fold size were the same as indicated in the results from the GridSearch below showing a max_df = .9, Regularization = 10, and using combination of unigrams and bigrams ((1,2)):

```
Results of cross-validation with 2 folds...
0.7689263141361534 {'classifier2__C': 10, 'features2__text__tfidf__max_df':
0.9, 'features2__text__tfidf__ngram_range': (1, 2)}
```

```
Results of cross-validation with 10 folds...
0.8039961465833194 {'classifier2__C': 10, 'features2__text__tfidf__max_df':
0.9, 'features2__text__tfidf__ngram_range': (1, 2)}
Results of cross-validation with 20 folds...
0.8066617017315212 {'classifier2__C': 10, 'features2__text__tfidf__max_df':
0.9, 'features2__text__tfidf__ngram_range': (1, 2)}
```

An analysis of adding new features was conducted by creating three new features representing the number of particles of speech (nouns, verbs and adjectives) in the text. The results showed that the accuracy remained the same and the recall even slightly decreased indication that these new features were not important for this classification task.

For each author, the model identified specific terms and text characteristics that were either highly relevant or less relevant for classification:

- Author 0 ("EAP"): Terms like 'mr', 'madame', and 'gentleman' were most indicative. The least important feature was 'idris'
- Author 1 ("HPL"): Terms included 'west', 'street', and 'later'. 'spirit' was the least important.
- Author 2 ("MWS"): Terms such as 'raymond' and 'perdita' were highly relevant

In the error analysis of the model's predictions, several patterns were observed that contributed to misclassifications. For instance, the model sometimes confused 'MWS' for 'HPL' when texts shared similar themes or when the narrative style was ambiguous. These misattributions were further complicated by high-weight features that were significant for one author but not for another, leading to confusion when such themes or styles overlapped. The numerical discrepancies in features such as number of adjectives or verbs led to errors, underscoring their inconsistent importance in different author styles. This suggests that the model may struggle with texts that exhibit shared stylistic features or thematic content, particularly when these elements are expressed through complex sentence structures or descriptions of similar settings.

The evaluation reveals that increasing the number of folds in cross-validation enhances model accuracy, suggesting better generalization but that after 10-folds the improvement diminishes significantly. Using strong regularization proved better than weak regularization strength. Using bi-grams and unigrams was more helpful than just unigrams. Ignoring a large number of common words appearing in 90%+ documents was better than words appearing in 95%+ documents. Feature importance analysis shows that specific keywords and stylistic elements significantly influence authorship classification. The error analysis highlights the need for further model tuning, particularly in better distinguishing shared stylistic features across different authors.