Compsci 361 Assignment 5

Hasnain Cheena  
190411106  
hche737

This assignment required the classification of research paper abstracts into their domains of research (Archaea, Bacteria, Eukaryota or Virus).

As the abstracts were plain text, a text pre-processing stage was required. Transformation of the abstracts was necessary to have them ready for the classification task. Therefore, a pre-processing stage was necessary. This stage consisted of tokenizing sentences into words, converting all words to lowercase and then removing punctuation, numbers and any stop words. A list of stop words was gathered from online (stop words list: <https://gist.github.com/sebleier/554280>).

Furthermore, to evauluate performance cross-validation using accuracy as a metric was used. 10-fold cross validation was performed with 20% holdout on the labelled dataset. The results were then averaged to get a reliable value for accuracy.

Note that the train-test split was completed before any text-preprocessing ensure that information was not shared from the test set to the training set.

Moreover, to evaluate performance improvements comparison to a baseline was necessary. Baseline performances were generated by using a majority class classifier and the standard Naive Bayes classifier. The majority class classifier had an accuracy of approximately 54% and standard Naive Bayes scored 93.4%.

*Extensions to Naive Bayes*

To improve the performance of the Naive Bayes two additions were made to standard Naive Bayes classifier. These include the implementation of the Multinomial Bayes classifier and using inverse document frequency instead of counts.

*1. Multinomial Bayes*

The feature input to the standard Bayes classifier are instances of boolean attributes, where each attribute is a 0 or 1 representing whether the word exists in the document. These boolean features were converted to frequency-based attributes where, each attribute is a count of how many times the word appeared in the specific document. This improved accuracy by 2.2% on average, taking the cross-validated accruracy from 93.4% to 95.2%.

*2. Inverse Document Frequency (IDF)*

The frequency -based attributes were then converted to inverse-document-frequency attributes. The idea behind the conversion is that commonly occurring words are unlikely to be correlated to a class and thus should be downweighted. In comparision, rarer words should be upweighted. Using inverse-document-frequencies increased the accuracy, on average, by 1.2% increasing it from 95.2 to 96.4%.

Therefore, a Bayes classifier using inverse-document-frequencies was decided as a good model for classification. The model was then retrained with all the avaliable data and used to predict an unlabelled dataset.