Compsci 361 Assignment 5

Hasnain Cheena
190411106
hche737

The task of this assignment is the classification of research paper abstracts into their problem domains.

As the abstracts were plain text, a text pre-processing stage was required. This processing stage transformed the abstracts into a suitable form for the Naïve Bayes classifier. The pre-processing consisted of tokenizing sentences into words, converting all words to lowercase and then removing punctuation, numbers and any stop words. The stop words were removed in accordance to a reference list (stop words list: https://gist.github.com/sebleier/554280).

Furthermore, to determine performance 10-fold cross-validation was performed with accuracy as the evaluation metric. The results were then averaged to get a reliable value for accuracy. Note that the train-test split was completed before any text pre-processing to 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 classifier two additions were made to the standard Naive Bayes classifier. These include the implementation of the Multinomial Bayes classifier and using inverse document frequency instead of counts.

1. Multinomial Bayes

To improve performance a Multinomial Bayes classifier was implemented. The feature input to the standard Naïve 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 1.8% on average, taking the cross-validated accuracy from 93.4% to 95.2%.

2. Inverse Document Frequency (IDF)

To further improve performance the frequency-based attributes were then converted to inverse-document-frequency attributes. The idea behind this conversion is that commonly occurring words are unlikely to be correlated to a class and thus should be down-weighted. In comparison, rarer words are more likely to correlate to a particular class and therefore should be up-weighted. Using inverse-document-frequencies increased the accuracy by 1.2%, on average, increasing it from 95.2 to 96.4%.

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