UPI: Jguo811
Name: Jianxing Guo

Implementation:

First step is reading the csv file and stores the data in an array. In order to use 10-fold-cross validation, we need split the data into 10 slices and each time use 1 slice to validate and 9 slices to train the model. In each training time, we have several steps to process data, train model and evaluate model. The data pre-processing step I did is data cleaning include remove stop words, remove numbers and lower all words. Then according the data after cleaning, create a dictionary that stores all unique words from the training data. Put each document in abstract into vector according the dictionary and save to matrix list. Then train the model, we need get the vector that is the probabilities of each word appear in each class in terms of the total dictionary. In the validation set, for each document from test set's abstract, we calculate the probabilities of each document in different class. The document should label with specific class that get the highest probability.

In my code, just run it and you can see the accuracy of 10-fold-cross validation of my standard multinomial Naïve Bayes and extensions classifiers.

The techniques I used to improve the classifiers include:

I use the frequency of each word to present the text data instead of Boolean representation, which can improve the classifiers. For data pre-processing, I clean up the data by removing the Stop Words and deleting numbers, which reduce the unnecessary features to improve the model. I use Laplace Smoothing to avoid computing problem because of the 0 value. Besides, when testing the new data, we may face some new words that did not in training dictionary and I also add smoothing in this step to avoid computing problem. I also use log technique to compute the probabilities instead of multiplying the probabilities, which can improve the accuracy.

Method extensions:

I also did some extensions of multinomial Naïve Bayes that is transforming by document frequency. It is a method that change the weight of probabilities of different words. If the word that occurs in most of instances, the weight should be decrease and if the word that occurs in few of instances, the weight should be increase because the word that most of instance include cannot help us to identify the class of instance.

Evaluation:

When we use the standard Naïve Bayes (Multinomial Naïve Bayes), the performance of 10-fold cross validation is about 93%. After implementing the extended Naïve Bayes method (transforming by document frequency), the performance of 10-fold cross validation is about 97%. About 4% of accuracy improves after implementing the extensions of Naïve Bayes that is transforming by document frequency because the weight of probabilities of words changes. If the words are important (low frequency) that can help us to identify the class, the weight of probabilities of this word should be increased. If the words are high frequency that almost all instances include this word that cannot help us to identify the class, the weight of probabilities of this word should be decreased.