Adjon Tahiraj
02/20/2019

MP1 Report

In my implementation of the bag of words model for multinomial, I did not neet the individual vectors voe each review, therefore I only returned a dictionary of the vocab words with the frequencies, along with the number of total words and the number of reviews needed for the multinomial implementation. For multinomial, I calculate the prior probabilities of each class positive and negative, then I calculate the probability of each word for each class and then compare the probabilities to see which prediction to make. For the bag of words for GaussianNB I returned a sparse matrix containing the frequencies of each word for each review, then I computed the mean and variance for each word and then used that to compute the GaussianNB classification of the test reviews. TO compute TF-IDF i computed the TF value first for each word and then I multiply that by the IDF value for each word to get a sparse matrix for each review containing the TF_IDF values. I also return the dictionary of the words in order to use them for the mean and variance calculations. I use these to compute the GaussianTFIDF classification on the test vectors.

I also used the stop words to remove them in the top of the extract clean words feature so none of the stop words are included in the vocab for any of the training or classification. I also changed the alpha for multinomial to 18 because that predicted the best accuracy.

Accuracy:
MultinomialNB_BagOfWords - 75.00%
GaussianNB_BagOfWords - 71.1 %
GaussianNB_TF_IDF - 49%

Running Time: 20 minutes