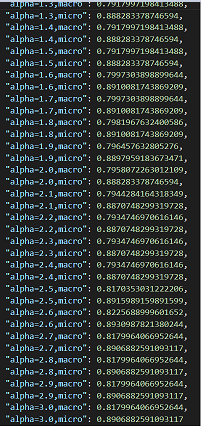
CMPE493 Assignment 2 Report

1. In pre-processing step following steps are performed:
   1. All news articles are extracted into a list.
   2. Topics are counted and top 10 most common topics are determined.
   3. News that has at least one of the most common topics are separated into training and test sets. On the development part of the algorithm, 5% of the training set was randomly taken as development set to tune alpha parameter.
   4. Prior probabilities and occurrences of each topic was calculated and saved.
   5. For multinomial Naïve Bayes approach, occurrence number of each word for each topic was determined and saved.
   6. For multivariate Bernoulli approach, occurrence fraction of each word in each class was calculated and saved.
   7. Total vocabulary of the dataset is determined and saved. Total size of the vocabulary is 19200 unique words.
   8. Test and development sets are saved for applications.
2. Top 10 topics are: earn, acq, money-fx, crude, grain, trade, interest, wheat, ship, corn. In the training set, occurrences of the top 10 topics are: 2877, 1650, 539, 391, 434, 369, 347, 212, 198, 183 respectively. In the test set, occurrences of the top 10 topics are: 1109, 798, 260, 241, 193, 183, 164, 94, 106, 70 respectively. 1523 of the documents have more than one topic.
3. Parameter tuning is conducted in following steps:
   1. 5% of the training set is randomly extracted to form a development set. The algorithm is then tested on the development set with alpha values starting from 0.1 and increasing it by 0.1 with each iteration. This tuning resulted in alpha=2.6 for multinomial Naïve Bayes and alpha=0.1 for multivariate Bernoulli.
4. Evaluation of the multinomial Naïve Bayes approach results in 0.840 precision, 0.879 recall and 0.858 f1 score in macro averaging; 0.907 precision, 0.936 recall, and 0.921 f1 score in micro averaging. Evaluation of the multivariate Bernoulli approach results in 0.513 precision, 0.943 recall, and 0.655 f1 score for macro averaging while it gives 0.581 precision, 0.953 recall, and 0.722 f1 score in micro averaging.

Text, calendar

Description automatically generated

Text

Description automatically generatedText

Description automatically generated