MACHINE LEARNING – CSE 6363

PROJECT 2

NAME: KARTHIK KUMARASUBRAMANIAN
STUDENT ID: 1001549999

SECTION OF PROBLEM:

The Problem can be divided into four sections,

- Data pre-processing.
- Splitting the data and label for test and train.
- Performing multinomial Naïve Bayes classification.
- Calculating the accuracy.

1. Data pre-processing:

The data is read from 20_newsgroups folder. The 20_newsgroups folder contains two sub folders Train and Test. Each subfolder contains 20 subfolders containing 500 files each. Each file is read line-by-line and split word-by-word into list of words for test as well as train. Each word from the list of words is checked for stopped words, punctuation, extra tabs and meta tag thereby reducing the number of words present in each list of words.

2. Test-Train split for Cross Validation:

From each list of words from train and test unique words and their count is calculated. With the word and count in each document a 2D array is created which is used as X_train and similarly for X_test. Each row of the X_train is mapped to respective folder class in y_train and y_test of same length as X_train and X_test.

3. Multinomial Naïve Bayes Classification:

The formula for Naïve Bayes Classification is given as,

$$h_{NB}(\mathbf{x}) = \arg\max_{y} P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

The probability of $P(x_i|y)$ is calculated as,

$$P(i|j) = \frac{word_{ij} + \alpha}{word_{j} + |V| + 1}, \alpha = 0.001$$

The log probability of the is used in-order to get better accuracy. The formula for log probability is given as,

$$Pr(j) = \log \pi_j + \sum_{i=1}^{|V|} f_i \log(t_i Pr(i|j))$$

4. Calculating the accuracy:

The classes and probabilities are predicted and calculated using the predict and log_probability functions respectively. The predicted classes and y_test lists are compared to get the accuracy of the model.

The accuracy from the model is given as 22.5 %. This may be due to the number of documents taken to train the model is same the number of documents to test the model. The model accuracy is because of underfitting and the number of documents to train the model needs to be more than the one to test it.

```
Accuracy is:  22.496749024707412
```

DATA:

The data for the problem is taken from the link, http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html. The data can also be found in the zipped folder under the name 20_newsgroups.zip.

REFERENCES:

1. https://towardsdatascience.com/multinomial-naive-bayes-classifier-for-text-analysis-python-8dd6825ece67 - For log probability calculation.
2. https://www.youtube.com/watch?v=EGKeC2S44Rs – For basic understanding of text classification using Naïve Bayes.