ML with Large Datasets:- ASSIGNMENT 1

Dinesh Kumar
dineshk@iisc.ac.in

1 Method Description

For predicting the class of a document the steps followed are:-

- In training set the no of occurrences of each word is counted, and output is a dictionary.

  $$\{word : \{class : number\}\}$$

- Number of occurrences of each class and the number of words in each class is counted and the output is again a dictionary.

  $$\{class : [Occurrence, words]\}$$

- Word count and test set requirement is combined for Final dictionary that will be constructed. output is in the form shown below.

  $$\{id, id - class : \{word : \{class : number\}\}\}$$

  where, id-class represents the class to which id be- longs to.

- The final,class count dictionary that was build will the basis for the final classification, and the algorithms for this are shown in the next sections.

2 Methods

the two different algorithms used are described.

2.1 Algo1:-

As was discussed in the class,this performs better when implemented on the Hadoop MapReduce Framework

$$logPr(y', x_1, x_2, .., x_d) = \sum_j log\frac{C(X = x_j \wedge Y = y') + mq_x}{C(X = ANY \wedge Y = y') + m} + log\frac{C(Y = y') + mq_x}{C(Y = ANY) + m}$$

Returns the best y'

2.2 Algo2:-

As mentioned in [1],this performs better when implemented on local machine.

$$logPr(y', x_1, x_2, .., x_d) = \sum_j log\frac{C(X = x_j \wedge Y = y')}{C(X = ANY \wedge Y = y') + m/q_x} + log\frac{C(Y = y') + mq_x}{C(Y = ANY)} \quad (2)$$

where, q x represents 1/(number of unique words in the training set)
Return the best y'

3 Architecture Used

1.`LocalImplementationtype1` : Here I have used a sim- ple python implementation of Naive Bayes classifier with- out using the above method sketch discussed.

2.`LocalImplementationtype2` : In this, using the above discussed methodology and ran the programs by using pipes.

3.`HadoopImplementation` : Used Hadoop MapReduce Framework to implement the the above discussed methodology. I have summarized the outcomes of the above methods in the table shown below.

method	Test	Validation	Training
LOCAL,A	24.89,33	30.3,64	48.35,180
LOCAL,B	70.09,33	69.86,61	69.43,180
LOCAL,C	13,3	13,4	10,5
CLUSTER,A	-	-	-
CLUSTER,B	-	-	-

Table 1: Classification accuracies, time(in minutes) for Naive Bayes on Local machine and cluster for various data sets.(A refers to Algo 1, B refers to Algo 2, C refers to type 2 implementation.)

4 References

1. https://www.3pillarglobal.com/insights/document-classification-using-multinomial-naive-bayes-classifier.

2. https://en.wikipedia.org/wiki/Amdahl