Assignment 1

V Nidhin Krishnan [1]

1. Model Description

There are two pass of map reduce for both train and test. One more map reduce pass is used to create a cache to be shared by all reducers. The cache is of size 3.5KB. For training in the first pass event counts like "Y=label X=word 22" are produced. In the next pass the output is such that the key is a word and values are count of word for a particular label. This output and test data are used as input for testing. In the first pass the output has the key as label of data and values as words and its counts. Then this output and test data are used to predict the labels of the test data. The final output is the id and its predicted labels, number of correct predictions and total number of data.

2. Code

The code is uploaded at https://github.com/nidhinkrishnanv/Naive-Bayes-classifier. Referred (Noll, 2011) for starter code for using python with hadoop .

3. Accuracy

	Train(%)	Development(%)	Test(%)
local	98.47	98.25	98.31
hadoop	97.91	73.44	77.01

4. Run time

	Train(sec)	Test(sec)
local	18.45	78.96
hadoop(10)	106	46
hadoop(1)	196	181

Runtime for local and hadoop run of map reduce. In the table hadoop(10) is the run time for 10 Reducers and hadoop(1) for 1 Reducer.

[1]Indian Institute of Science, Bangalore. Correspondence to: V Nidhin Krishnan <nidhinkrishnanv@gmail.com>.

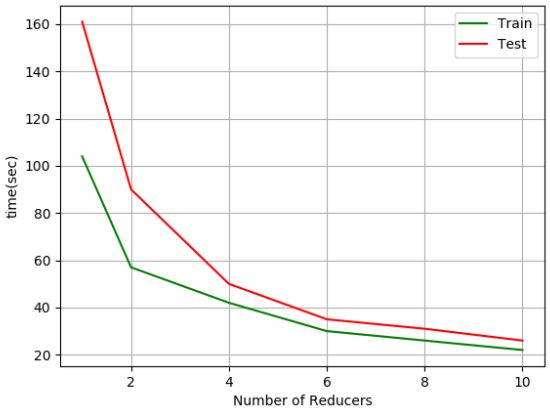

Figure 1. Run time for number of reducers from 2 to 10

The wall time is not linear with respect to number of reducers. From 2 to 4 reducer there is exponential decrease in time. Then it starts to flatten out from 8 to 10. Which can be see in Figure 1.

5. Number of parameters

	Parametes
local	2232301
hadoop	2230872

References

Noll, Michael G. http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/, 2011.