Naive Bayes Classification using PIG

William Cohen

Due Wed, April 23, 2014 13:29pm EST via Autolab
50% credit if submit before Fri, April 25, 2014 13:29pm

Policy on Collaboration among Students

These policies are the same as were used in Dr. Rosenfeld's previous version of 10601 from 2013. The purpose of student collaboration is to facilitate learning, not to circumvent it. Studying the material in groups is strongly encouraged. It is also allowed to seek help from other students in understanding the material needed to solve a particular homework problem, provided no written notes are shared, or are taken at that time, and provided learning is facilitated, not circumvented. The actual solution must be done by each student alone, and the student should be ready to reproduce their solution upon request. The presence or absence of any form of help or collaboration, whether given or received, must be explicitly stated and disclosed in full by all involved, on the first page of their assignment. Specifically, each assignment solution must start by answering the following questions in the report:

- Did you receive any help whatsoever from anyone in solving this assignment? Yes / No. If you answered 'yes', give full details: _____ (e.g. "Jane explained to me what is asked in Question 3.4")
- Did you give any help whatsoever to anyone in solving this assignment? Yes / No. If you answered 'yes', give full details: _____ (e.g. "I pointed Joe to section 2.3 to help him with Question 2".

Collaboration without full disclosure will be handled severely, in compliance with CMU's Policy on Cheating and Plagiarism. As a related point,

some of the homework assignments used in this class may have been used in prior versions of this class, or in classes at other institutions. Avoiding the use of heavily tested assignments will detract from the main purpose of these assignments, which is to reinforce the material and stimulate thinking. Because some of these assignments may have been used before, solutions to them may be (or may have been) available online, or from other people. It is explicitly forbidden to use any such sources, or to consult people who have solved these problems before. You must solve the homework assignments completely on your own. I will mostly rely on your wisdom and honor to follow this rule, but if a violation is detected it will be dealt with harshly. Collaboration with other students who are currently taking the class is allowed, but only under the conditions stated below.

1 Assignment

In this assignment, you are required to write a pig script which takes word event counts and test documents and outputs a list of 1000 best documents for a target class based on log likelihood scores. Instead of classifying a document against all target classes and picking the best class, we are going to classify each document against a single target class and get a list of best documents where the list is sorted from highest to lowest log likelihood scores.

2 PIG

Pig is a high level language for creating a chain of MapReduce jobs. Instead of writing java program using MapReduce API, one uses Pig Latin. Pig Latin is a language similar to SQL where the user can use different kind of operators to load data from HDFS, perform operations like filtering, grouping, joining etc (similar to SQL) and finally store final results back to HDFS. For more information, see `http://pig.apache.org/docs/r0.11.1/basic.html`. In order to help you, we are listing the steps required to accomplish this in Pig.

3 Steps

The pig script can be broken down in to the following steps:

1. load the data from HDFS

2. Use raw word events to obtain marginal counts for words (word events)
3. Use raw word events to obtain joint word, label counts (joint events)
4. Use word events to obtain vocabulary size (needed for smoothing)
5. Use word events, joint events and vocabulary size to compute the score of each word for given target class i.e $log(P(w/target))$
6. Use the previous output and test documents to compute likelihood score for test documents without using the class prior i.e $\sum_{w \in d} log(P(w/target))$ (log likelihood)
7. Use raw word events to obtain class prior for each class (class prior)
8. Use log likelihood and target class prior to compute total log likelihood
9. Use the previous output and test documents to obtain the final output

Each step in the above description corresponds to few Pig Latin statements. you will need to use operators like JOIN, FILTER, CROSS, GROUP etc.

4 Running PIG script on AWS EMR

Running pig script on AWS EMR is very similar to running a streaming job. When you create your cluster do not remove PIG from the list of extra softwares apart from Hadoop. Now you will have an option to select Pig program from the drop down menu in Steps. you can then configure and add. you will need to specify the s3 path to your pig script, input and output. We have already uploaded the words events and test documents to s3 so you will not need to do so. To load the input from within your pig script, you can use $INPUT, $OUTPUT. you can either specify input / output in the space provided or pass it through command line using -p flag.

5 Input & Output Format

The word event counts file has the following format:

Each line is of the form: $word = w, lab = l < TAB > count$ OR $word =$

$w < TAB > count$ OR $lab = l < TAB > count$

The test documents file has the following format:

Each line is of the form: $docID < TAB > true\_labels < TAB > bag\_of\_words$

Your final output should have the following format:

$docID < TAB > true\_labels < TAB > log\_likelihood\_score$

The target label for the output is ECAT.

6 Data

The data for this assignment is already in the following s3 buckets:
The word event counts are in s3://bigmldatasets/rcv1_train_events/.
The test documents are in s3://bigmldatasets/rcv1_test/full/ &
s3://bigmldatasets/rcv1_test/small/.

7 Deliverables

You are required to submit the logs from successful run of your pig script on AWS EMR. Apart from this, you are also required to submit your final output. Note that you just need to submit a single tar file **hw7.tar** that include all the files above to Autolab, and the log file and output will be verified manually by the TA, and no immediate feedback will be provided for this assignment.

8 Marking breakdown

- you will receive full points if you are able to successfully run your pig script on AWS EMR: this will be verified by the correctness of your log file and your final output.