MIE 1624 Introduction to Data Science and Analytics – Winter 2017

Assignment 2

Due date: Thursday, March 23 at 11:59pm

Background

There has been growing interest, in recent years, in mining online political sentiment in order to predict the outcome of elections. Continuing on from Assignment 1, you are asked to implement more sophisticated machine learning approaches and apply them to calculating the sentiment of tweets, in particular the sentiment of tweets related to the 2015 Canadian federal election.

As in Assignment 1, you have access to a classified data set, *classified_tweets.txt*, and an unclassified data set, *unclassified_tweets.txt.* The classified data set contains tweets that have had their sentiments already analyzed and recorded as values 0 and 4. A value of 0 is a negative tweet and a value of 4 is a positive tweet.

You are asked to implement two of the following well established approaches: logistic regression, linear regression, SVM, and Naive Bayes, select one of those two based on analyzing their performance on the classified data set and use it to categorize the tweets in the unclassified data set. You are also asked to implement an algorithm to categorize the tweets in the unclassified data set by political party.

Produce a report detailing the analysis you performed in order to choose the most suitable classifier between the two initial selections, the results of applying the chosen algorithm to the unclassified tweets, as well as any potential insights into the political sentiment of the Canadian electorate with respect to the major political parties participating in the 2015 federal election.

Finding the most suitable algorithm for a given application task (comparing multiple classifiers on a specific domain) requires selecting performance measures, such as *accuracy, true positive rate* (TPR), *false positive rate* (FPR), etc., according to which to judge the algorithms being considered. As part of your evaluation procedure, select an appropriate number of measures, review the results that your chosen methods obtain on the selected measures and try to explain these observations. Among the questions to ask here is: "Can the observed results be attributed to the characteristics of the implemented classifiers or are they observed by chance?" Hypothesis testing is one way to gather additional evidence of the extent to which the results of the evaluation metrics are representative of the general behaviour of the classifiers under consideration. As the data we have access to is limited (and it is not the whole population of tweets), statistical re-sampling techniques (simple or multiple resampling, etc.) can be applied on the data available in order to improve the estimation of the classification error. (Keep in mind that, given sufficient data, it is always possible to show that a difference between two alternatives, no matter how small, is statistically significant.) When evaluating the performance of two algorithms, also keep in mind that there can be an inherent tradeoff between the results on various performance measures. For example, the TPR and the FPR are quite different and often an algorithm with good results on one yields bad results on the other. Classifier evaluation can also be viewed as a problem of analyzing high-dimensional data and various methods can be employed for an effective visualization. In your report, present at least one graphical comparison of the

performance of the classification algorithms you have selected.

Learning Objectives

(1) Understand how to apply machine learning algorithms (e.g., Logistic regression, Linear regression, SVM, and Naive Bayes) to the task of text classification.

(2) Improve on skills and competencies required to compare and contrast the performance of classification algorithms on one domain (text classification), including application of performance measurements, statistical hypothesis testing and visualization of comparisons.

(3) Improve on skills and competencies required to collate and present domain specific, evidence-based insights.

To Do:

- Apply two of the following algorithms: Logistic regression, Linear regression, SVM, Naive Bayes, to the task of classifying tweets into positive and negative tweets
- Prepare a 3 to 4-page report describing:
  - (1) the classification algorithms you have implemented (at most 1.5 pages);
  - (2) the results of applying the two selected algorithms on the classified dataset, according to the selected performance measures, as well as your interpretation of the results, including graphical representations of the comparisons between the chosen classifiers, and your choice of method to apply to the unclassified data set;
  - (3) the results of applying the chosen algorithm on the unclassified dataset, as well as any potential insights into the political sentiment of the Canadian electorate with respect to the major political parties participating in the 2015 federal election.

All graphs should be readable and have all axes appropriately labelled. All visual materials should be understandable and all graphs should be appropriately labelled and easy to read.

Tools Required

- ***Software***
  - **Python Version 3.X Only NO** 2.7 is allowed for this assignment. Your code should run on the Data Scientist Workbench (Kernel 3). All libraries and builtins are allowed but here is a list of the major libraries you might consider: Numpy, Scipy, Scikit, Matplotlib, Pandas.
- ***Data files***
  - **corpus.txt:** file containing a set of words and associated sentiment value
  - **stop_words.txt:** file containing an extensive list of stop words
  - **classified_tweets.txt:** file containing a set of tweets which have already been classified as negative (have a sentiment score of 0) or positive tweets (have a sentiment score of 4)

<u>What to submit via BlackBoard:</u>

(1) IPython notebook and an equivalent Python .py file containing your implementation of the classifiers and the various evaluation methods.
   (a) **lastname_studentnumber_assignment2.ipynb**
   (b) **lastname_studentnumber_assignment2.py**

(2) Your four-page report named **lastname_studentnumber_assignment2.pdf**

Respect the above convention when naming your files, making sure that all letters are lowercase and hyphens are used as shown. Only submissions via Blackboard will be accepted.

BEFORE uploading your files, make sure that:

(1) your file name **does not** contain any extras, such as version information, e.g., **lastname_firstname_studentnumber_assignment2.py**.

(2) you comment your code appropriately and describe your algorithms in sufficient detail.