PROEA 821-005 Spring 2018 Machine Learning: Twitter polluter Detection

1 Introduction

Each example in this classification task is a Twitter user. The goal is to predict whether the user is a content polluter or not.

2 Data

The data-splits directory contains the following three data files (one training set and two test sets) and three index files:

1. `data-splits/data.train`: This is the training set, in the usual libSVM format. There are 29049 training examples.
2. `data-splits/data.test`: This is the set of examples on which you will report results in your final report. There are 6224 test examples.
3. `data-splits/data.eval.anon`: These 6226 examples are all labeled positive in the provided data set. You should use your models to make predictions on each example and upload them to Kaggle. See below for the format of the upload. Half of these examples are used to produce the public leader board. The other half will be used to evaluate your results.
4. `data-splits/data.train.id`: This index file for training set. Each row contains a user id corresponding to each example in `data.train`.
5. `data-splits/data.test.id`: This index file for test set. Each row contains a user id corresponding to each example in `data.test`.
6. `data-splits/data.eval.anon.id`: This index file for the evaluation set. Each row contains a user id corresponding to each example in

`data.eval.anon`.In addition, the ids from this file will be used to match your uploaded predictions on Kaggle.

In all, there are 16 features. Note that as part of your project, you are welcome to try feature space expansions, kernels, other non-linear methods or extract fetures by yourself.

3 Tweet

Besides the features, we also provide the original dataset which can help you extract new features. The raw-data directory contains the following three data files:

1. `raw-data/all-users.txt`: This file contains the user information in the form of "`UserID\tCreatedAt\tCollectedAt\tNumerOfFolloings\tNumberOfFollowers\tNumberOfTweets\tLengthOfScreenName\tLengthOfDescriptionInUserProfile`"

2. `raw-data/followings.txt`: This file contains the following information in the form of "`UserID\tSeriesOfNumberOfFollowings`"(Each number of following is separated by ,)

3. `raw-data/tweets.txt`: This file contains the tweets information in the form of "`UserID\tTweetID\tTweet\tCreatedAt`".

4 Evaluation

We will use the accuracy score to evaluate the classifiers. The examples are all split randomly among the three files. So we expect that the cross-validation performance on the training set and the F1 scores on the test set and the public and private splits of the evaluation set will be similar.

5 Submission format

Kaggle accepts a csv file with your predictions on the examples in the evaluation data. There should be a header line containing `example_id,label`. Each subsequent line should consist of two entries: The example id (from the file `data.eval.ids`) and the prediction (0 or 1).

We have provided two sample solutions for your reference:

1. `sample-solutions/sample-solutions.all.positive.csv`: Where all examples are labeled as positive

2. `sample-solutions/sample-solutions.half-neg.csv`: Where the first half of examples are labeled false and the second half are labeled true