CIS 520, Machine Learning, Fall 2017: Final Project
Multi-class Classification on Tweets

1 Overview

For the final project, you will be given tweets, and your job is to correctly classify them as joy/sadness/surprise/fear/anger. This is a multi-class classification task, and the results will be evaluated with corresponding costs. The training dataset given consists of a collection of 18092 labeled samples. Your classifier will be tested on a validation (or "quiz") set for the leaderboard, and a hold-out test set for final ranking. The features of the dataset are described in more detail in the README.txt file of the project kit. The format of the project is a competition, with live leaderboards (see below for more details).

2 Project Rules and Requirements

Rules and Policies

- You **CANNOT** download or harvest any additional training data from the internet. Both of these will be considered as cheating, and the penalty will be very harsh. Please, **ONLY** use the data we provide. We will test your final classifier, and if it is clear that we cannot replicate your performance because you use additional data, you will get a **ZERO** for the project.
- Except when specified otherwise, you are allowed to download additional code or toolboxes from the internet. However, you must cite everything you use in your final project report. We do not want you to reinvent the wheel. If you are unsure of what extra resources are allowed, please ask us.
- You must work in groups of 2 or 3 people. **No single competitors or groups with more than 3 people** will be allowed.
- Before you can submit to the leaderboard, you need to **register your team** using turnin (described below).
- You will need to ensure that your code will run from start to finish on our server, so that we can reproduce your result for grading. We will provide a utility so that you can make sure we can run your code successfully. See below for details.

Overall requirements

The project is broken down into a series of checkpoints. There are four mandatory checkpoints and a final writeup which is due Dec.12th. The leaderboard will be operating continuously so you can monitor your progress against other teams and towards the score based checkpoints. All mandatory deadlines are midnight. So, the deadline Nov.18th means you can submit anytime before the 18th becomes the 19th.

- 1% - Nov.22, run turnin -c cis520 -p proj_groups group.txt to let us know your team name. The file should contain a single line: the team name. In order to post scores to the leaderboard, you must have submitted your team name and have 2-3 members of the team total.
- 9% - Nov.28, Checkpoint: Beat the baseline 1 by any margin.
- 20% - Dec.4, Checkpoint: Beat the baseline 2 by any margin.
- 50% - Dec.8, By the final submission (Dec 8 11:59 PM), implement (or download and adapt an implementation of) the following 3 methods as well as your own model. Please submit your implemented models to 'proj_final' on turnin. We should be able to immediately and manually run both training and testing for each of your three algorithms easily. (Note: turnin will not automatically run your code or email you an output for your review.) Please include a README or PDF briefly describing the models you trained and documenting which files correspond to each algorithm and instructions on how to run and test with them. (You do not need to submit the final report at this time.). From these for or more models submitted to proj_final, they do not necessarily have to include the model used for your leaderboard, but you may! We will have your leaderboard model submission in the turnin project submission outlined below.

  NOTE: There are **TWO** submissions expected on Dec. 8th. ONE is turnin to **proj_final** and ANOTHER is turnin to **project**. The **project** submission is elaborated in Section 5 to verify that your leaderboard-worthy model satisfies the timing constraints and correctness.

  1. A generative method (NB, HMMs, k-means clustering, GMMs, etc.)
  2. A discriminative method (logistic regression, decision trees, SVMs, etc.)
  3. An instance based method (kernel regression, k-nearest neighbors, etc.)
- 20% - Dec.12 Submit the final report as a PDF to Canvas by 11:59 PM on December 12. The final report should be between 2 and 5 pages and should include all of the following.

  1. Results for each method you tried (try to use checkpoints to get validation set accuracy for each of your methods)

  2. Analysis of your experiments. What worked and what did not work? Why not? What did you do to try to fix it? Simply saying "I tried XX and it did not work" is not enough. Give an explanation.
- Extra credit - In the competition, placing well will increase your project grade. First place gets 10%, second 8%, third 7%, and the rest of the top 10 teams 5% extra added to the project grade.

3 Evaluation

In this project, instead of simple precision, we will consider the relation between labels to define the cost when predicting a wrong label. The cost matrix is as follows:

	joy	sadness	surprise	anger	fear
joy	0	3	1	2	3
sadness	4	0	2	3	2
surprise	1	2	0	2	2
anger	2	1	2	0	2
fear	2	2	2	1	0

Note that the first column represents the ground truth, and the first row is the predict labels. Take the first row as an example, if the ground truth is 'joy' and your predicted label is 'sadness', then the cost is 3. If your predicted label is 'surprise', the cost is 1. Your classifier should make the cost as **low** as possible.

4 Requirements for Each Checkpoint

For the second and third checkpoints, you must submit to the leaderboard(s). For the final checkpoint, you must submit ALL of your code via turnin to the correct project folder. Make sure that you submit any code that you used in any way to train and evaluate your method. We will be opening up an autograder that will check the validity of your code to ensure that we will be able to evaluate it at the end.

5 Detailed Instructions

You will be submitting your code to the auto-grader, which will execute your code on the validation set and generate a vector of predictions. The auto-grader then will compare the predictions with the ground truth. You will be receiving an email enclosing the cost. The cost will be recorded on the leaderboard.

Register your team name

Before you can get results on the leaderboard, you need to submit your team name. Everyone on your team is required to do this. Simply create a text file on eniac with your team name as follows:

$ echo 'My Team Name' > group.txt

$ turnin -c cis520 -p proj_groups group.txt

This group.txt file should be raw text and contain only a single line. Do not submit PDFs, word documents, rich text, HTML, or anything like that. Just follow the above commands. If you have a SEAS email address, then you will get an email confirmation.

Submit to the leaderboard

Submit all your code files including predict_labels.m and all supporting files needed (models):

turnin -c cis520 -p leaderboard

Your team can submit **once every 3 hours**, so use your submissions wisely. Your submission will be checked against the reference solutions and you will get your score back via email. This score will also be posted to the leaderboard so everyone can see how awesome you are.

You can view the current leaderboard here: leadboard fall 17

Submit your code for the final checkpoint or to test correctness

Your code will be predict_labels.m which takes as arguments the feature files (as in the sample code from the kit). The time constraint for making predictions on 9,098 validation samples is **10 minutes**.

You must submit your code for the final checkpoint. You can do so with the following:

turnin -c cis520 -p project ⟨list of files including predict_labels.m ⟩

You will receive feedback from the autograder, exactly like the homework. The feedback you will get from the autograder is whether your prediction code runs within 10 minutes for 9,310 test samples and whether the submission size is less than 50 Mb. You will not get feedback about your error on the validation set,

which appears on the leaderboard. The final rankings will be released on the day of the prize ceremony, Dec. 12.