

Overview

- This miniproject is a competition hosted on Kaggle. You can use any open-source tools and MATLAB, using both concepts you learned in class as well as any other techniques you find online, to get the best out-of-sample error rate that you can.
- This miniproject is due 5pm, February 11th, via Moodle. You can work in groups of up to three, but must make submissions from a single account.
- At the end of the miniproject, in addition to submitting predictions online you must submit a report including documented code that goes through every iteration of your model, explaining your thoughts and results.

A Machine Learning Competition in Sentiment Analysis

In this competition, you are going to predict the sentiment, either opposition or support, expressed by a speech.

You will be provided with a bag of words representation of each speech. The words are the 1000 most common words in the speeches, stemmed, and with stop words filtered out. 'training_data.txt' contains 4189 speeches you can use to train your model. The first row contains the heading of the 1000 words, followed by the label heading. Each subsequent row contains the count of each word in the given speech, followed by a label indicating the sentiment of the speech. We labeled '1' on the report if the speech indicated support, and a '0' on the report if the speech indicated opposition.

'testing_data.txt' contains 1355 testing points, in the same format, but excluding the label. You will submit prediction files in the format of 'sample_solution.txt'. Each row in your prediction file should have an id, and a prediction. The first prediction should have id 1, and the id increments each following row, in the same order as in testing_data.txt. The prediction will be a label, 0 or 1.

Your evaluation on the test set will be classification accuracy, what percentage of points you labeled correctly. A public leaderboard will display results accuracy on roughly 50% of the predictions in testing_data.txt. When the competition is over, your best entry will be ranked on a private leaderboard, comprising of the remaining roughly 50% of test points not included in the public leaderboard. You will be evaluated on both your public and private leaderboard performance.