Machine Learning Final Report

Sentiment Analysis of Movie Reviews

By: Keaton Tagseth,

Janakiram,

and Vivek Kumar

Introduction: Presently, lot of companies depend on Machine techniques to improve or manage their products, investments, profit, losses based on the reviews which they get from Customers. But processing the reviews is tedious task to come up to conclusion. The reviews could be a mixture of Alphabets, numbers, Internet-Slang words, Smiley's, .gif and .png images which people express their feeling to a particular item. Converting each expression to some form of meaning and checking if it is positive or negative is simple for one review, but having lots of dataset like 12MB. So, we have used sentiment analysis methods to find the reviews are positive or negative.

Background: We planned work on the topic apart from assignments. First we worked on Spam Classifier using Support Vector Machines Algorithm(linear kernel) from Andrew Course Lectures. We observed the flow of Dataset being processed like dataset cleaning, Extracting the features, Indices and final output. After we build list of words and expression for Movies Review's project.

Approach: We came into our project with a massive amount of information on the subject, including many ways of solving our classification issue. After reading extensive amounts of papers on sentiment analysis, we came to the conclusion of using a support vector machine algorithm. We researched and read up on many other algorithm types such as linear regression and a naïve bayes, but with the size of our dataset, a support vector machine algorithm was agreed upon. Linear regression was thrown out pretty much instantly as our massive amount of data would create so many features that a linear regression algorithm wouldn’t be very efficient. We ended up using a support vector machine because of there being a large section of the coursera dedicated to going over them. After following through the coursera lectures on SVM’s, we went through the exercise 6 problem set also from the coursera course. This problem set was almost identical to the problem sets that we had in class with the exception of no written math problems at the and. Once this problem set was completed, we heavily modified it to work with our movie reviews instead of the spam e-mails that were detailed in the exercise. This included adding in support for more words; including internet slang and smiley ‘emojis’, creating new vectors and matrices for our data set, refining the kernel algorithm and how our movie reviews were processed. Initially we had planned to implement a stemming algorithm, which ended up working and correctly stemmed the dataset, but we

implemented it too late to make use of it. We would have had to remake our matlab environment and our bag of words which we could not have done in the time allotted.

Dataset: We used a set of data comprised of 25000 labelled movie reviews from the website imdb.com. From this set of data, we ended up creating a word indices matrix of the first 5000. Initially we had it set to learn the first 12000 reviews, but when it was finished matlab gave an error while saving the environment and we had to run it again, but did not have enough time as learning the 12000 took over 14 hours to run. We also had a cross-validation set of 1000 reviews which gave us an accuracy of 79%. The final part of the data was the unlabeled set of 11000 reviews which we classified with an accuracy of 82%. I ran the program again after the presentation and competition on kaggle had completed with our program trained with the 12000 reviews, and the accuracy improved to 85%.

1. Sample negative movie review: Caught this on IFC yesterday, and can't believe the positive reviews! Am I the only one who thought these "ladies" were anything but? Kate tells Jed she could get fired because she's supposed to be a pillar of the community, but puts out for him! Then they suddenly decide they're in love? And she's SO devastated over his death, she doesn't go to his funeral, much less, tell his family the "good news"! By the way, how did an American get to be the headmistress of a very proper British school? Janine should have been kicked off the force for her inexcusable abuse of power, but nothing happens! And she winds up boffing a con she brought in for questioning! And the less said about Molly, the better!

As for the guilt Janine and Molly feel over Jed, please! It's the punk's own damn fault he got turned into roadkill! Where's the guilt over poor Gerald, who gets puked on? If only I could do the same to the bozos behind this "movie"!

Evaluation: Evaluating all of the movie reviews took an incredible amount of time to complete. Once our algorithm was completed, in order to analyze the dataset, we have to fill a word indices matrix which took about an hour for every 1000 reviews. This could have been sped up if we reduced the amount of words in our bag of words and if we made use of our stemming algorithm. The first thing we did was create our list of words in a bag of words format. After analyzing every review of the 25000 labeled reviews, we found around 80,000 unique words, which doesn’t help us a whole lot. After creating our bag of words, we created our word indices matrix which corresponded to our bag of words. Once we trained our machine with the first 2500 negative reviews and 2500 positive reviews, we found which words were most significant which included, perfect; alien; enjoyed; excellent; and favorite. Once we had our dataset evaluated, we had to train our SVM which we used a simplified version of a sequential minimal optimization algorithm, which was perfect for what we were doing. We then ran our set of 11000 unlabeled movie reviews through our machine and storing the results in a comma separated value document. The results ended up being higher than our cross validation set, which is abnormal and we ended up getting an accuracy rating of around 82%.

Conclusion: Processing a dataset and giving the review is straight forward. Apart from movie dataset, we observed some of reviews are in the form .img and .png for Instance facebook. In future, processing this images dataset needed advanced algorithms.

Team Roles: As a team we did large amounts of research on the subject in order to make informed decisions on our project such as what kind of algorithm to implement. Janakiram implemented the stemming algorithm and added support for all of the 'emojis' which our program possibly had to handle. Vivek Kumar was in charge of tokenizing the data. Keaton Tagseth did the bulk of the programming, including the completion and modification of the original exercise 6 from Andrew Ng's coursera course.

References:

http://mlwave.com/movie-review-sentiment-analysis-with-vowpal-wabbit/

http://theanalysisofdata.com/gl/sentimentWorkshop.pdf

https://cs224d.stanford.edu/reports/TimmarajuAditya.pdf

https://www.coursera.org/learn/machine-learning