Jeremy Mednik, Jiayi Du, Stephen Owen
3-8-20

Project 1 Report

Our general penalized regression does not do a good job at classifying extreme and moderate categories. The training set is 79 percent accurate and the validation set is only 74 percent accurate. The base model generates 335 coefficients corresponding to the best lambda value. To better visualize it, lambda equal to 25 was chosen. words like “terrible,” “horrible,” “decent,” and “nice” which are often used to describe writers’ emotion or evaluation of the movies could be viewed as indicators to predict the two categories. However, these words are not accurate enough to predict the category of the review because there are some limitations. These strong words often used to describe a certain aspect of the movie like the story or music rather than the whole movie. In addition, a contrast word may be used after the sentence containing adjectives, so the tone of the review could be overturned even though a few strong adjectives appear in the previous sentences.

Selecting only some part-of-speech tags for our penalized regression model gives only slightly lower classification rates; 79% on the training set and 73% on the validation set. Some of the features selected by this model include lots of adjectives and adverbs, ex: “well”, “much”, “too”, “little”, “ever”, “pretty”. This makes sense because these words can tell us what the reviewer thinks of the movie being reviewed. For example: “The movie is well-acted by all, …” shows that this reviewer likes something about the movie. The word “too” in the following review shows to us that this review is more negative than positive: “a tendency to use too much unnecessary dialogue…”. Our model makes many errors. An extreme review may be classified as moderate if the word “well” is in it, and a moderate review may be classified as extreme if it has a word like “terrible” in it even if that word is used to describe something other than the movie being reviewed.

After constructing a model that focused on certain parts of speech, we directed our attention to a model that used bi-grams, or n-grams that can include two sequential lemmas as a coefficient value for the model. We first decided to make a type token ratio: the average number of unique words divided by the number of all words. The number of unique words for extreme reviews was nearly the same as that for moderate reviews (55.3% and 54.2% respectfully). Since there were a lot of repeating words compared to the amazon dataset, we wanted to see which combinations of words carried value in predicting the category. We yielded an accuracy rating of 74.8% which is around the same as the other models. Some of the bi-grams coefficients we saw were “not even” and “ever see,” proving to be insightful when we saw the lemmas in context. Lastly for this model, we wanted to see which reviews the model felt most confident in predicting. Most of them were correct, but the most confident one was actually incorrect, mistaking an extreme review for a moderate one, but that makes sense because there was no strong feeling of hate from the review.

We then thought an interesting angle to take would be to focus on the examples that the model misclassified, and see if we could gather any sort of insight into the types of things that are holding our previous models back. We created a number of individual models that used exclusively the training examples that were misclassified by our prior models, all of which performed about the same as the models trained on the entire dataset (even with 10, 20, and 30 cross-validation folds). This leads us to believe that with the types of models we’ve studied this year, the theoretical maximum accuracy attainable is close to what our models have achieved (around 80% accuracy on the training set, and slightly less on the validation set). However, analyzing the misclassified examples did lead to some

interesting quirks. For example, the word “ever” consistently showed up in our prior models as being heavily associated with extreme reviews. However, when models were run on the misclassified data, “ever” became heavily associated with moderate reviews. The same inversion is true of the word “pretty” (although the change went from moderate to extreme). Another example of these quirks was that the words “but” and “little” ended up becoming heavily weighted towards extreme reviews when in actuality they weren’t used in any indicative contexts.