Assignment 3 Report - Naive Bayes Classifier

Mark Tedder

November 29, 2018

1 Data Sorting

Each text file and broke it apart into 84 reviews for training and left the remaining 10 positive and 11 negative reviews for testing. A complete positive and a complete negative review list was created in order to use all the reviews to train the classifier for the test set.

2 Vocabulary

Initially 2 vocab dictionaries were constructed, one for each class, positive and negative. In these dictionaries all the words and their respective counts were collected from each review. I broke apart each review string by words and punctuation, as well as lowering the case of all the words in order to avoid multiple entries for the same word. While adding each word from each review to the dictionary I kept a counter in the for loop to keep track of the total number of words in both the positive and negative reviews, necessary for later calculations. Later I constructed these same dictionaries using all the reviews, keeping the same overall word counts.

3 Training Naive Bayes

Using the number of reviews in the positive and negative reviews, as well as their combined total, I create a dictionary of the log-likelihood of being positive or negative overall. From the vocab dictionary from above, I obtain counts of each word, the total number of different words, as well as the the total number of words in the dictionary, adjusted by Laplace smoothing to assign each word in the vocab a log-likelihood in a dictionary. The log-likelihood being the log of the probability calculated below in (1), where $|V|$ is the number of individual words in the vocabulary.

$$\mathbb{P}(word|class) = \frac{count(word|class) + 1}{\sum_i count(word_i|class) + |V|} \tag{1}$$

To account for words that are found in the test case that did not appear in the training data, they are assigned the probability of a word that appears once in the given class vocabulary. This probability is kept in the log-likelihood dictionary corresponding to the word "UNK", again being the log of the probability below.

$$\mathbb{P}(unknown) = \frac{2}{\sum_i count(word_i|class) + |V|} \tag{2}$$

4 Testing Naive Bayes

To test the viability of our collected log-likelihood probabilities as well as the class probabilities, the log-likelihood of each word in each the test review along with the log-likelihood of a review being a given class are summed. We see this below in (3), where $|R|$ is the total number of words in the review.

$$\log(\mathbb{P}(class|review)) = \log(\mathbb{P}(class)) + \sum_{i=1}^{|R|} \log(\mathbb{P}(word_i|class)) \tag{3}$$

If a given word in the review isn't in the vocab dictionary of a given class, the probability of the word "UNK" is added to the sum. Each class, positive and negative, not have a corresponding probability, and the greater of the two is what is given as the classification for the test review.

5 Naive Bayes Classifier

A text file of id numbers and reviews can be taken and a classification of either positive or negative next to the corresponding id number can be returned. The reviews are separated each assigned a "POS" or "NEG" classification, then the id numbers and their corresponding classifications are concatenated and exported in a text file.