Christian Zinck
11/20/18

1.

The methods I used to preprocess the data were the bag of words and word embedding. For the bag of words method, I chose to use the word count vector variant as opposed to the binary variant because I wanted to be able to represent when a reviewer felt so strongly about a word that they used it twice, but it runs the risk of misidentifying double negatives. I set the minimum frequency to 1/1000, so if a word only appeared once in the dataset, meaning it was likely a misspelling, I ignored it. Additionally, I set the maximum frequency to 1/10, so words like “the”, “and”, “you”, “if” and “so” were ignored as well.

The other method was a word embedding. For my reference embedding, I used the GloVe embedding provided to us. To convert a sentence to a vector, I did element-wise addition of all the embeddings of the words in the sentence, as was recommended.

In terms of general data processing, I chose to leave apostrophes in the words and remove all other punctuation because I wanted to maintain contractions. I also processed the data in a case insensitive way to ensure that I would lump all instances of the same correctly spelled word together. By doing this I sacrificed my the ability to gleam the intensity of a word due to capitalization.

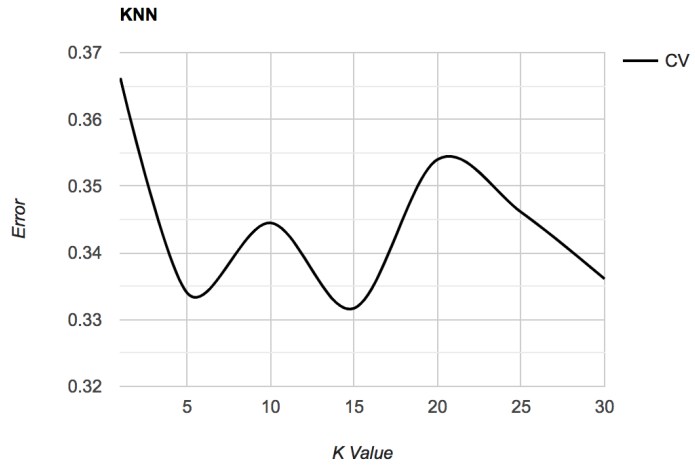
2.

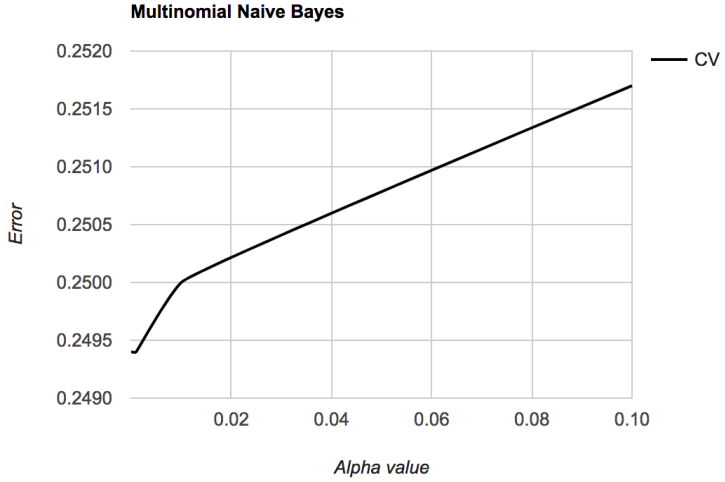
The three classification algorithms I used were KNN, Multinomial Naive Bayes, and Logistic Regression. For KNN, the hyperparameter is obviously the number of nearest neighbors k. I selected the range [1, 30] to test and found that 15 maximized CV accuracy for both training sets. For MNB the hyperparameter is an additive smoothing factor alpha. I selected the range (0, 1] to test (the entire parameter range) and found that a low, but nonzero value of alpha maximized CV accuracy, so I chose .001, as alpha values closer to zero stopped producing different results. For Logistic Regression the hyper parameter is the inverse of regularizer strength C, meaning that a small C value will cause high regularization and a large C value will cause low regularization. I selected the range [.001, 1000] to test and found that a C value of 1 gave the best CV results.

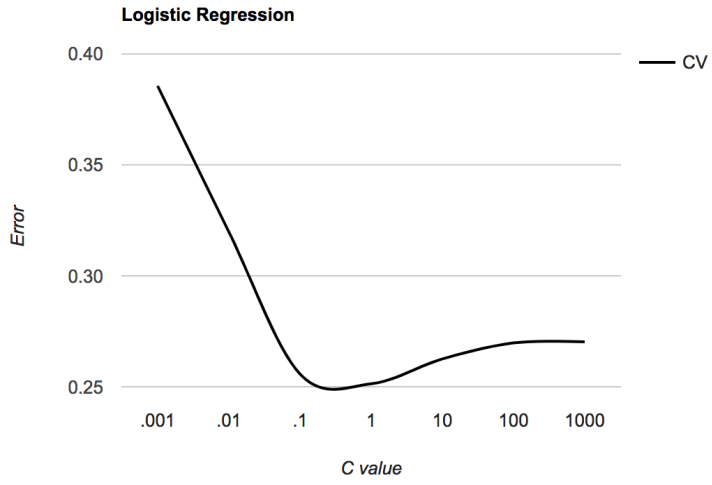
3.

Since I made all the hyperparameters arguments to my classification script, I was able to do all of my model selection from the command line. I simply tested many different values within the range and found which ones produced consistently better CV results.

4.


KNN
CV
0.37
0.36
0.35
0.34
0.33
0.32
Error
5
10
15
20
25
30
K Value


Multinomial Naive Bayes
CV
0.2520
0.2515
0.2510
0.2505
0.2500
0.2495
0.2490
Error
0.02
0.04
0.06
0.08
0.10
Alpha value


Logistic Regression
CV
0.40
0.35
0.30
0.25
Error
.001
.01
.1
1
10
100
1000
C value

5.

Again, the hyperparameter values were k = 15, a = .001, and c = 1.

Bag of Words -

KNN:	[0.66575491, 0.67079709]
MND:	[0.74922152, 0.75196848]
Logistic Regression:	[0.74593571, 0.75086029]

Word Embedding -

KNN:	[0.66835646, 0.67169354]
Logistic Regression:	[0.75235737, 0.75445063]

Furthermore, due to the majority vote I did, I expect the accuracy confidence interval for my submitted predictions to be significantly higher than any of these individual intervals, and I expect the interval itself to be much narrower.