HOMEWORK 3

(g) The train error by training the data contained in XTrain and YTrain is 0

The test error by training the data contained in Xtrain and Ytrain is 0.020690 = 2%

By looking at the above two errors we found that there is no significant difference in the error of the training set and the test error. This is because the algorithm is tested on the same documents, which is being used to train the classifier. There are no new records that the classifier has seen during testing, so error is zero in case of train.

(h) The train error by training the data contained in XTrainSmall and YTrainSmall is 0.096552=9.6%

The test error by training the data contained in XtrainSmall and YtrainSmall is 0.27586=27%

By looking at the above two errors we found that there is significant difference in the error of the training set and the test error.

When we have less training data, the prior have more impact on our classifier. As, β values are given here and we have taken less data compared to training data, the β values will have some impact on the error.

But on the other hand, taken very large amount of data in consideration, as n increases values of beta will not be considered impactful and prior will not add any information to our analysis. High values of α will compensate for the effect of the β values.

(i) For each class label, the five words that model says are most likely to occur in a document from class y are :

List 1 :

Class 2

Word1 =a

Word 2=and

Word3=the

Word4=to

Word 5=of

Word6= in

Class 1

Word1 =the

Word 2=to

Word3=of

Word4=in

Word 5=a

List 2

Top words :

when Y=1 and Y ≠1

Vocabulary(find(v==words2(1,i)), 1)

organis

reckon

favour

centr

labour 1990s

when Y=2 and Y≠2

4enlarg

5enlarg

monday

percent

realiz

The list of words describes the two classes better is list 2. In the list 1, all the words we get are either prepositions or conjunctions that are mostly found in all the documents irrespective of the topic.

But in list 2, all the words we get are specific to the context of the topic 'The economist' and 'the onion'. By looking at these words we can say that the word might have come from the specific documents but in list 1, looking at the words , it is difficult to say which document does it belong to.