Assignment 3

Yingnan Zhao 260563769

Question 2

set the baseline using the random classifier and the majority classifier

```
Most Frequent F1 Score= 0.182384900074
Uniform Random F1 Score= 0.212874083355
```

Decision Tree

A grid search has been done to find the best combinations of the three hyperparameters. For the maximum number of leaf feature I considered the range from 300 to 100, the maximum feature parameter I considered from 1 to 0.5 the min sample split I considered from 0.1 to 0.00097

```
the optimal number of leafs is 200, min sample split is:1.0, max features is
:1.0, it achieves a F1 score of 0.380983427942
```

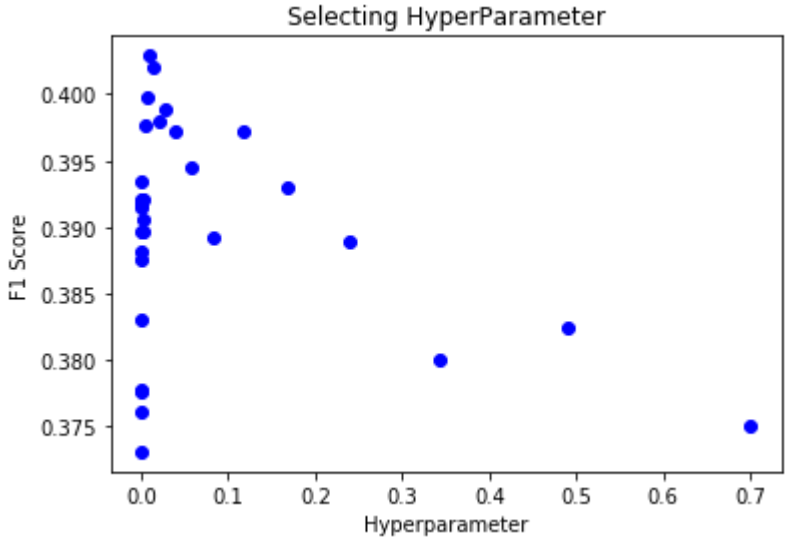
Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of decision Tree for yelp train set frequency bag of word repre
sentation is 0.51882181684
The F1 score of decision Tree for yelp valid set frequency bag of word repre
sentation is 0.380648384418
The F1 score of decision Tree for yelp test set frequency bag of word repres
entation is 0.3832236497
```

Bernoulli Native Bayes

The graph below shows the process of choosing the best hyperparameters for the Bernoulli Native Bayes classifier to classify the yelp binary bag of word dataset, there is one hyper parameter to be adjusted, which is Additive (Laplace/Lidstone) smoothing parameter, aplha.

The optimal alpha will be find in the range of 0.7 to 0.0000225.




```
the optimal alpha is 0.009688901040699992, it achieves a F1 score of 0.40289
32560404406
```

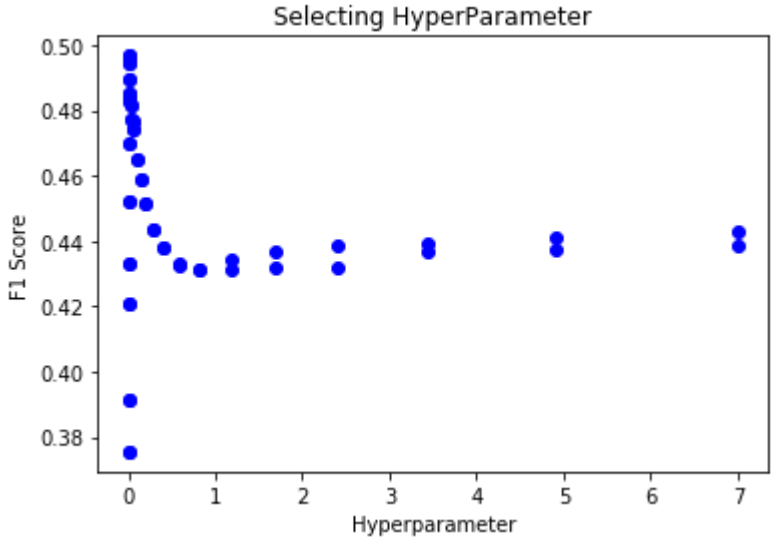
Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of Native Bayes for yelp train set binary bag of word represent
ation is 0.740081113991
The F1 score of Native Bayes for yelp valid set binary bag of word represent
ation is 0.40289325604
The F1 score of Native Bayes for yelp test set binary bag of word representa
tion is 0.43113819783
```

Linear SVM

The graph below shows the process of choosing the best hyperparameters for the linear SVM to classify the data, there are two hyperparameters to be considered, C and dual. C is the penalty parameter of the error term, and dual is to select to either to solve the dual or primal optimization problem.

Finding the best combination of the two hyperparameters. C is considered from 7 to 0.000225, dual is either True or False




```
the optimal C is 0.007979226629761192, dual= True, it achieves a F1 score of
0.4967606230498824
```

Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of LinearSVM for yelp train set binary bag of word representati
on is 0.833265595284
The F1 score of LinearSVM for yelp valid set binary bag of word representati
on is 0.49676062305
The F1 score of LinearSVM for yelp test set binary bag of word representatio
n is 0.501717243121
```

Comment about the performance of different classifers.

For the binary bag of word representation of the yelp set, Linear SVM Classifier gives the best result in all three datasets (train, valid, test), and decision gives the worst result, however without setting the hyperparameters to make the decision tree classifier more general, the training F1 score is 1. The linear SVM classifier gives much better result than the other two classifiers. Generally, all three classifier produce result much better than the baseline.

Question 3

Decision tree

A grid search has been done to find the best combinations of the three hyperparameters. For the maximum number of leaf feature I considered the range from 300 to 100, the maximum feature parameter I considered from .95 to 0.5 the min sample split I considered from 0.005 to 0.0000097

```
the optimal number of leafs is 110, min sample split is:0.01, max features i
s :1.0, it achieves a F1 score of 0.373826114467
```

Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of decision Tree for yelp train set frequency bag of word repre
sentation is 0.4834558775156819
The F1 score of decision Tree for yelp valid set frequency bag of word repre
sentation is 0.37382611446735314
The F1 score of decision Tree for yelp test set frequency bag of word repres
entation is 0.38287396168355314
```

Gaussian Native Bayes

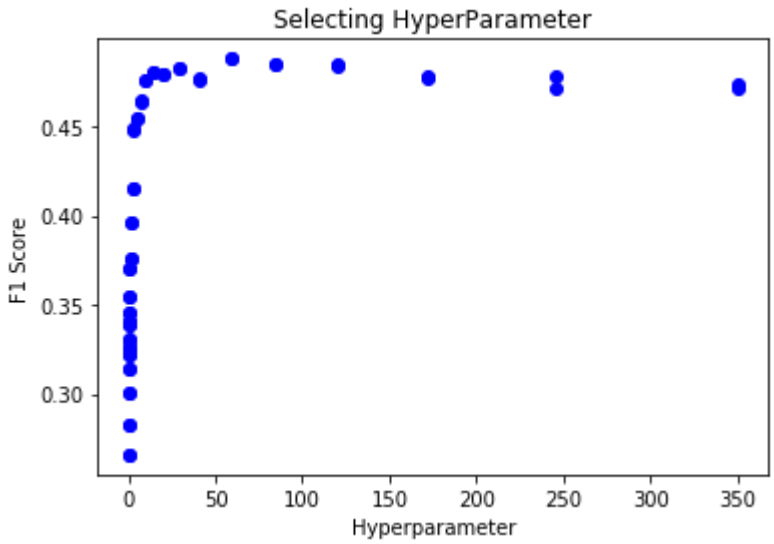
The result below shows the result of using Gaussian Native Bayes classifier to classify the data, there is no hyper parameter to be adjusted.

```
Using Gaussian Native Bayes on yelp train set, F1 score: 0.807218481283
Using Gaussian Native Bayes on yelp valid set, F1 score: 0.299617796185
Using Gaussian Native Bayes on yelp test set, F1 score: 0.312702436953
```

Linear SVM

Linear SVM classifier have been applied to the yelp freq bag of word, the graph below is showing the process of finding the best combination of the two hyperparameters, C and dual. C is considered from 350 to 0.55, dual is either Ture or False.

```
the optimal C is 58.82449999999997, dual= True, it achieves a F1 score of 0.
4880968952693054
```



Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of LinearSVM for yelp train set frequency bag of word represent
ation is 0.829556100128
The F1 score of LinearSVM for yelp valid set frequency bag of word represent
ation is 0.488096895269
The F1 score of LinearSVM for yelp test set frequency bag of word representa
tion is 0.498574911495
```

Comment on the result and compare with last question

Linear SVM still perform the best, however the Decision tree classifier performs much better than Gaussian native bayes for this dataset. There is a big drop in performance in Native Bayes, as there is no hyperparameter to be adjusted, and the probability distribution might not be Gaussian. And the performance of the other two classifiers remain stable. So there is no improvement in proformance using the frequency bag of word representation.

I think for language processing the word appears most often are the words that has little meaning like "the", "a", "I". And in the frequency bag of word representation they weighted a lot, althought the classifiers should be able to tell those words have little effect on the result through training, they still affected the result. If those words are removed, we should be able to get better result. Also the order of word matters.

Question4

set the baseline using the random classifier and the majority classifier

```
Uniform Random F1 Score= 0.497523170342
```

Decision tree

A grid search has been done to find the best combinations of the three hyperparameters. For the maximum number of leaf feature I considered the range from 300 to 100, the maximum feature parameter I considered from .95 to 0.5 the min sample split I considered from 0.5 to 0.0000097

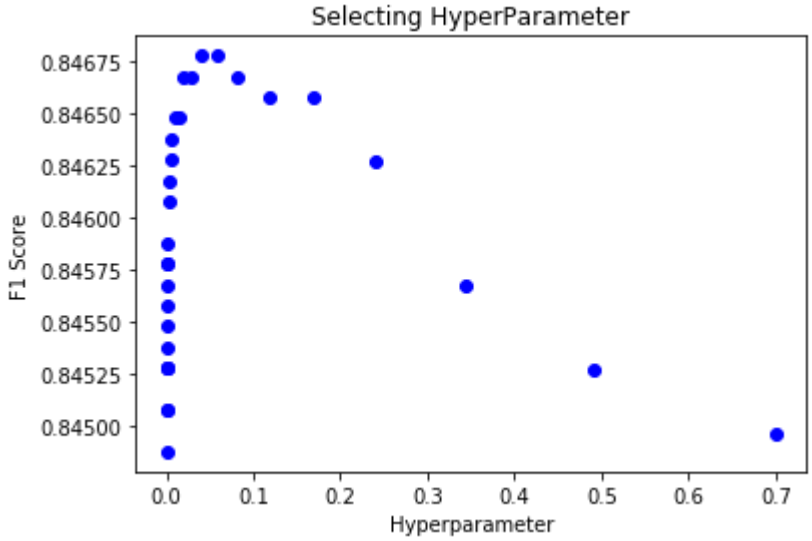
```
the optimal number of leafs is 120, min sample split is:0.5, max features is
:1.0, it achieves a F1 score of 0.749424405219
```

Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of decision Tree for IMDB train set binary bag of word represen
tation is 0.792050102249
The F1 score of decision Tree for IMDB valid set binary bag of word represen
tation is 0.749424405219
The F1 score of decision Tree for IMDB test set binary bag of word represent
ation is 0.748565486964
```

Bernoulli Native Bayes

The graph below shows the process of finding the optimal hyperparameter for the the Bernoulli Native Bayes classifier, there is one hyper parameter to be adjusted, which is Additive (Laplace/Lidstone) smoothing parameter, aplha. Alpha will be find in the range of 0.7 to 0.0000225.



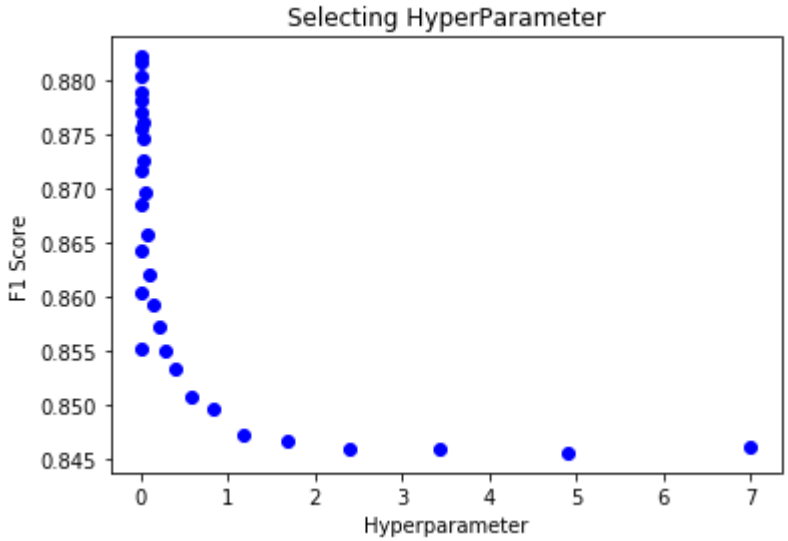

```
the optimal alpha is 0.04035360699999998, it achieves a F1 score of 0.846774
1048237153
```

Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of Bernoulli Native Bayes for IMDB train set binary bag of word
representation is 0.871388945268
The F1 score of Bernoulli Native Bayes for IMDB valid set binary bag of word
representation is 0.844782168186
The F1 score of Bernoulli Native Bayes for IMDB test set binary bag of word
representation is 0.828050189863
```

Linear SVM

The process of finding the best penalty parameter of the error term C is shown in the graph below, dual parameter is not considered because according to the documents, dual = False is prefered when the number of samples is larger than the number of features.




```
the optimal C is 0.005585458640832834, it achieves a F1 score of 0.882190028
5640178
```

Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of LinearSVM for IMDB train set frequency bag of word represent
ation is 0.953360313143
The F1 score of LinearSVM for IMDB valid set frequency bag of word represent
ation is 0.883273880301
The F1 score of LinearSVM for IMDB test set frequency bag of word representa
tion is 0.877204130262
```

comments

For the IMDB binary bag of word dataset, the linear SVM still perform the best, and the decision tree still perform the worst, all three classifiers are still performing much better than the base line.

In the section below, optimal hyperparameters will be found for the IMDB data sets in Frequency bag of word representation.

Decision Tree

A grid search has been done to find the best combinations of the three hyperparameters. For the maximum number of leaf feature I considered the range from 300 to 100, the maximum feature parameter I considered from .95 to 0.5 the min sample split I considered from 0.5 to 0.0000097

```
the optimal number of leafs is 300, min sample split is:0.0625, max features
is :0.95, it achieves a F1 score of 0.7467424316645439
```

Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of decision Tree for IMDB train set Frequency bag of word repre
sentation is 0.7647251523291618
The F1 score of decision Tree for IMDB valid set Frequency bag of word repre
sentation is 0.7467424316645439
The F1 score of decision Tree for IMDB test set Frequency bag of word repres
entation is 0.75
```

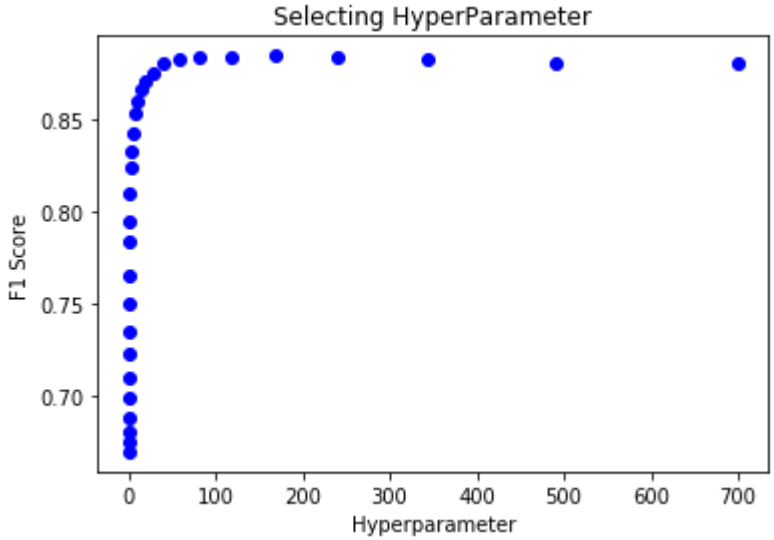
Gaussian Native Bayes

For Gaussian Native Bayes Classifier there is no hyperparameter to be adjusted.

```
Using Gaussian Native Bayes on IMDB train set, F1 score: 0.847282881077
Using Gaussian Native Bayes on IMDB valid set, F1 score: 0.736549363568
Using Gaussian Native Bayes on IMDB test set, F1 score: 0.638794006467
```

Linear SVM

The process of finding the best penalty parameter of the error term C is shown in the graph below, dual parameter is not considered because according to the documents, dual = False is prefered when the number of samples is larger than the number of features.




```
the optimal C is 168.06999999999994, it achieves a F1 score of 0.88429649996
91241
```

Applying the optimal hyperparameters learnt to the test set.

```
The F1 score of LinearSVM for IMDB train set frequency bag of word represent
ation is 0.960559626915
The F1 score of LinearSVM for IMDB valid set frequency bag of word represent
ation is 0.884932869219
The F1 score of LinearSVM for IMDB test set frequency bag of word representa
tion is 0.875890213651
```

comments

The SVM classifier still performs the best and decision tree the second. performance of the Native Bayes deops the same way as using the yelp data, when using frequency bag of words representation for IMDB datasets. Generally the IMDB dataset have better performance than yelp dataset, this is mainly because it only have 2 possible outcomes compares to 5. When switching between two ways of representing the data, the yelp and IMDB datasets follows the same pattern.