COMP-551: Applied Machine Learning - Assignment #3

Antonios Valkanas

Student ID: 260672034

Question 1

Please see the submitted folder for this question.

Question 2

a) As a baseline, for the purposes of testing our model we used two dummy classifiers. The first one was a uniform-class classifier and the second one was a majority-class classifier. The following results were observed:

```
Random Dummy Classifier F1: 0.185
Majority Dummy Classifier F1: 0.356
```

b) Next, Naïve Bayes, Decision Tree and Linear SVM classifiers were to be used. However, to properly use these classifiers it was necessary to tune their hyperparameters. To achieve this, I used sci-kit learn and Parameter Grid for various parameters on the validation set.

c) The results of tuning the parameters are summarized as follows:

- Parameter tuning for Naïve Bayes: α-values tested: [1e-5, 1e-4, 1e-3, 0.01, 0.1, 0.3, 0.6, 1].
  From the tests the optimal value was determined to be α = 0.01, with F1 = 0.428.
- Parameter tuning for Decision Tree: max depth: [None,5,10,100,500], min. samples split: [3,5,10,15,20], criterion: [entropy, Gini]
  From the tests the optimal values for the hyperparameters are: max depth = 10, min. sample split = 10, criterion = entropy, with F1 = 0.413.
- Parameter tuning for Linear SVM: C: [0.5,1.0,1.5,2.0,5.0,10.0,100.0], loss: [hinge, squared hinge]
  From the tests the optimal values for the hyperparameters are: C = 0.5, loss = squared hinge, with F1 = 0.465.

Note: For the full test results please consult the submitted folder for this question.

d) The following F1 resulted for the tuned models:

```
Optimal NB f1 for train, validation, test: (0.7478571428571429, 0.428, 0.4395)
Optimal Tree f1 for train, validation, test: (0.5022857142857143, 0.413, 0.3895)
Optimal SVM f1 for train, validation, test: (0.9931428571428571, 0.465, 0.4475)
```

e) The performance on the test set did not point to an obviously superior method. However, the SVM was slightly better than the Naïve Bayes and quite a bit better that the Decision Tree. The most important parameter for the SVM is C which was chosen to be 0.5.

Question 3

a) Next, Naïve Bayes, Decision Tree and Linear SVM classifiers were to be used. However, to properly use these classifiers it was necessary to tune their hyperparameters. To achieve this, I used sci-kit learn and Parameter Grid for various parameters on the validation set.

b) The results of tuning the parameters are summarized as follows:

- Parameter tuning for Naïve Bayes: α-values tested: [1e-5, 1e-4, 1e-3, 0.01, 0.1, 0.3, 0.6, 1].

From the tests the optimal value was determined to be α = 0.01, with F1 = 0.425.

- Parameter tuning for Decision Tree: max depth: [None,5,10,100,500], min. samples split: [3,5,10,15,20], criterion: [entropy, Gini]
  From the tests the optimal values for the hyperparameters are: max depth = 5, min. sample split = 5, criterion = entropy, with F1 = 0.400.
- Parameter tuning for Linear SVM: C: [0.5,1.0,1.5,2.0,5.0,10.0,100.0], loss: [hinge, squared hinge]
  From the tests the optimal values for the hyperparameters are: C = 100.0, loss = squared hinge, with F1 = 0.491.

Note: For the full test results please consult the submitted folder for this question.

c) The following F1 resulted for the tuned models:

```
Optimal NB f1 for train, validation, test (0.7482857142857143, 0.425, 0.437)
Optimal Tree f1 for train, validation, test (0.5434285714285715, 0.39, 0.3785)
Optimal SVM f1 for train, validation, test (0.48328571428571426, 0.431, 0.4545)
```

d) The performance on the test set did not point to an obviously superior method. However, the SVM was slightly better than the Decision Tree and quite a bit better that the Naïve Bayes. The most important parameter for the SVM is C which was chosen to be 100.

e) We observe: Naïve Bayes and Decision Tree decrease in performance, which is probably due to noise in the data (as we do not remove common words such as 'the', 'a' etc. which throws the frequency analysis off. On the other hand, SVM performed slightly better for the same reason.

f) Using the FBoW is expected to improve the performance of the model. The reason why FBoW performs better than BBoW is because when certain words are used more than once such as 'bad' we can get a better idea for what the review rating will be than if all the information we have is whether the word is used or not. To make full use of frequency analysis we need to remove commonly used words that do not give us insight into the review from the bag of words as discussed in the previous part.

Question 4

a) As a baseline, for the purposes of testing our model we used two dummy classifiers. The first one was a uniform-class classifier and the second one was a majority-class classifier. The following results were observed:

```
Random Dummy Classifier F1: 0.223
Majority Dummy Classifier F1: 0.356
```

b) Next, Naïve Bayes, Decision Tree and Linear SVM classifiers were to be used. However, to properly use these classifiers it was necessary to tune their hyperparameters. To achieve this, I used sci-kit learn and Parameter Grid for various parameters on the validation set.

c) The results of tuning the parameters are summarized as follows:

- Parameter tuning for Naïve Bayes: α-values tested: [1e-5, 1e-4, 1e-3, 0.01, 0.1, 0.3, 0.6, 1].
  From the tests the optimal value was determined to be α = 0.01, with F1 = 0.841.
- Parameter tuning for Decision Tree: max depth: [None,5,10,100,500], min. samples split: [3,5,10,15,20], criterion: [entropy, Gini]
  From the tests the optimal values for the hyperparameters are: max depth = 10, min. sample split = 10, criterion = entropy, with F1 = 0.736.
- Parameter tuning for Linear SVM: C: [0.5,1.0,1.5,2.0,5.0,10.0,100.0], loss: [hinge, squared hinge]
  From the tests the optimal values for the hyperparameters are: C = 0.5, loss = squared hinge, with F1 = 0.847.

Note: For the full test results please consult the submitted folder for this question.

d) The following F1 resulted for the tuned models:

```
Optimal NB f1 for train, valid, test (0.872, 0.8423615337796714, 0.8318656900666611)
Optimal Tree f1 for train, valid, test (0.771, 0.7296233839235526, 0.7262984336356142)
Optimal SVM f1 for train, valid, test (1.0, 0.8468016843793864, 0.836321341194401)
```

e) The performance on the test set did not point to an obviously superior method. However, the SVM was slightly better than the Naïve Bayes and quite a bit better that the Decision Tree. The most important parameter for the SVM is C which was chosen to be 0.5.

Question 5

a) Next, Naïve Bayes, Decision Tree and Linear SVM classifiers were to be used. However, to properly use these classifiers it was necessary to tune their hyperparameters. To achieve this, I used sci-kit learn and Parameter Grid for various parameters on the validation set.

b) The results of tuning the parameters are summarized as follows:

- Parameter tuning for Naïve Bayes: α-values tested: [1e-5, 1e-4, 1e-3, 0.01, 0.1, 0.3, 0.6, 1].
  From the tests the optimal value was determined to be α = 0.6, with F1 = 0.840.
- Parameter tuning for Decision Tree: max depth: [None,5,10,100,500], min. samples split: [3,5,10,15,20], criterion: [entropy, Gini]
  From the tests the optimal values for the hyperparameters are: max depth = 10, min. sample split = 10, criterion = Gini, with F1 = 0.744.
- Parameter tuning for Linear SVM: C: [0.5,1.0,1.5,2.0,5.0,10.0,100.0], loss: [hinge, squared hinge]
  From the tests the optimal values for the hyperparameters are: C = 100.0, loss = squared hinge, with F1 = 0.878.

Note: For the full test results please consult the submitted folder for this question.

c) The following F1 resulted for the tuned models:

```
Optimal NB f1 for train, valid, test (0.870, 0.8404352689921692, 0.8298329750289399)
Optimal Tree f1 for train, valid, test (0.795, 0.7427874139283535, 0.7464284472870103)
Optimal SVM f1 for train, valid, test (0.950, 0.8784158415841584, 0.874476820664089)
```

d) The performance on the test set did not point to an obviously superior method. However, the SVM was slightly better than Naïve Bayes and quite a bit better that the Decision Tree. The most important parameter for the SVM is C which was chosen to be 100.

e) We observe: Naïve Bayes decreases in performance, which is probably due to noise in the data (as we do not remove common words such as ‘the’, ‘a’ etc. which throws the frequency analysis off. On the other hand, SVM and Decision Tree performed slightly better for the same reason.

f) Using the FBoW is expected to improve the performance of the model. The reason why FBoW performs better than BBoW is because when certain words are used more than once such as ‘bad’ we can get a better idea for what the review rating will be than if all the information we have is whether the word is used or not. To make full use of frequency analysis we need to remove commonly used words that do not give us insight into the review from the bag of words as discussed in the previous part.

g) From the results, it is clear that the IMDB dataset yielded much better results than the Yelp dataset. The reason for this is that while the features are similar, the IMDB dataset is a binary problem while the Yelp dataset has 5 possible outputs for each set of features. As a result, when the model is not sure about the answer in the IMDB case (classification probability is close to 50%) is has a good chance of guessing the answer, while in the Yelp case it could in theory be a guess with each class having a 20% probability of being chosen. Furthermore, it is possible that a yelp review is given a 4 instead of a 5 when the model recognizes that the review is positive but fails to find how positive the comment was. On the other

hand, broadly classifying the review as either positive or negative is a simpler problem which is why the IMDB dataset gave us better results for the same methods.