Homework 1-Report

Mamoun Almardini (800845973)

In this report we discuss the results achieved by building two text classifiers (Naïve Bayes and Decision Tree) on the 20newsgroups dataset. The report is divide into three parts; the first part presents the processing applied on the data before analyzing it. The second two parts discuss the results of the Naïve Bayes classifier and Decision Tree classifier respectively. In this assignment, I used Matlab Machine Learning Toolbox and I organized my implementation, as explained in the README.txt, to match the structure of the files described in the assignment.

1. **Data Reprocessing:**

The data is presented in a bag-of-words format; which shows the document-ID, word-ID and frequency of the words. In order to work on the data, we need to present the data in a different format; where the columns represent the words IDs and the rows are actually the documents objects. I used the script provided by the assignment, with few modifications, to transform the dataset into the needed format.

2. **Naïve Bayes (NB) Classifier:**

- Accuracy: following is the accuracy achieved after the cross validation and after applying the test data on the trained model:
  - Cross validation: 85.62% (using 10-fold cross-validation)
  - Test Data= 78.11%

As we can see here, the accuracy achieved against the test data is less than the accuracy achieved from the cross validation, which is logical; given the fact that in the test data we may encounter examples that were not present in the train data.

- The Effect of Regularization Parameter (alpha):

In order to find the change in the cross-validation as alpha changes, I had to write a script that loops over the prior probabilities calculated in the training model and update them based on alpha value. Figure 1 shows the effect of changing the regularization parameter on the cross validation accuracy.

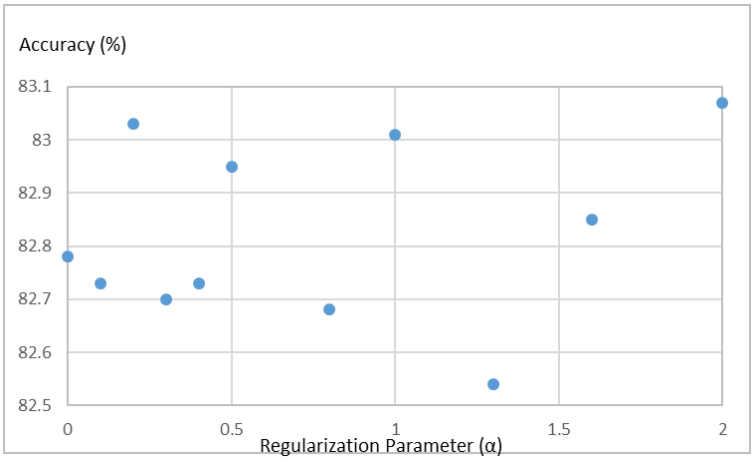

Figure 1 Regularization Parameter (α) Effect on the Model's Accuracy

As seen in Figure 1, the accuracy for the values within the range 0-0.5 are centered around 82.8%, which is close to the accuracy measured by matlab without changing alpha. Please note that I used 3-fold for the cross validation.

- Running Time:

    - Training: 1 min
    - Cross-Validation (10-fold): 13 min
    - Testing: 7 min

3. Decision Tree (DT) Classifier

- Accuracy: following is the accuracy achieved after the cross validation and after applying the test data on the model:

    - Root of the tree is word ID: 16628
    - Cross validation: 58.97% (using the full tree)
    - Test Data= 50.74%

- Pruning the Tree: Table 1 shows the effect of pruning the tree on the prediction accuracy. I have found that the best depth of the tree is 35.

Table 1 The effect of decision tree pruning on the accuracy

Depth	Cross Validation Accuracy (%)	Accuracy Against the Test Data (%)
10	58.67	52.19
35 (best pruning depth)	58.6	53.3
40	58.65	53.07
Full depth	58.97	50.74

As seen in Table 1, the accuracy achieved from the test data is lower than the accuracy achieved from the cross validation, which is logical, as explain earlier in the Naïve Bayes results analysis. In addition to that, note that the accuracy achieved from cross validation using the full tree is the best among others. On the contrary, the accuracy of the full tree against the test data is not the best. The reason is due to the fact that the tree keeps growing deep enough until it fits the training data (overfitting). Therefore, pruning the tree would allow us to have a more generalized tree that performs better with the unseen data.

- Running Time:
    - Training: 10 min
    - Cross-Validation (10-fold): 70 min
    - Testing: 30 min

Conclusion

In this assignment, we applied Naïve Bayes and Decision Tree classifiers on the 20newsgroup dataset. We presented the results of each classifier and showed the difference between the calculated accuracies. Finally, we found that Naïve Bayes classifier performed better than Decision Tree on the given dataset.

*Running the code: please refer to the README.txt for a full description of the files and how to run the code.