GE461 - Introduction to Data Science

Project 5 - Data Stream Mining

Instructor: Prof. Dr. Fazlı Can

Mustafa Yaşar
Bilkent University, 06800. Çankaya/Ankara

Question 1

In this question, hyperplane datasets have been generated by using the hyperplane generator which is provided by skmultiflow library. The generated dataset contains 20,000 samples and each sample contains 10 features and a binary label.

- **a.** The first dataset has a noise percentage of 10 and number of drifting features equal to 2. The generated dataset has been written into a file called "Hyperplane Dataset 10_2"
- **b.** The second dataset has a noise percentage of 30 and number of drifting features equal to 2.
- **c.** The third dataset has a noise percentage of 10 and number of drifting features equal to 5.
- **d.** The fourth and last dataset has a noise percentage of 30 and number of drifting features equal to 5.

Question 2

In this question, different online classifiers which are trained and tested by using the datasets generated in the questions are created. There are hyperparameters such as the maximum sample size in HT and maximum window size in KNN for the online classifiers. By using different hyperparameters, optimized parameters for the classifiers were determined.

- **a.** The hyper parameter for the Hoeffding Tree classifier is the maximum sample size. The best accuracies have been obtained when the maximum sample size is 5000 for the first dataset, 20,000 for the second dataset, 20,000 for the third dataset, and 1000 for the fourth dataset.

The best accuracies are 84.8%, 63.2%, 80.1%, and 65.5% for the datasets, respectively.

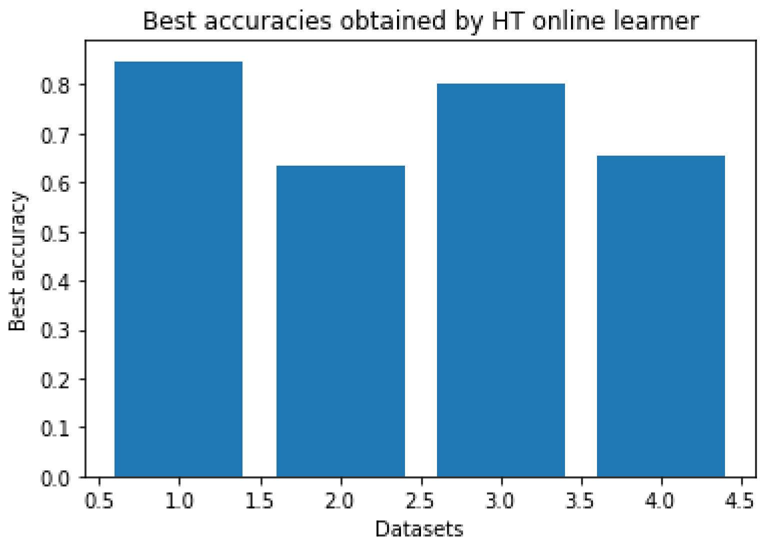

Figure 1. Best accuracies obtained by HT classifier

The figure above shows the accuracies obtained by using the datasets generated in the first question.

b. The hyperparameters that should be determined for the KNN classifier are maximum window size and the number of neighbors.

The best accuracy was obtained for the first dataset when the maximum window size is 10,000 and the number of neighbors is 15.

Similarly, the best accuracy was obtained for the second dataset when the maximum window size is 10,000 and the number of neighbors is 15.

The best accuracy was obtained for the third dataset when the maximum window size is 5000, and the number of neighbors is 15.

The best accuracy was obtained for the last dataset when the maximum window size is 5000, and the number of neighbors i s15.

Best accuracies when hyperparameters are like above are: 82.1%, 63.04%, 83.5%, and 62.7% respectively.

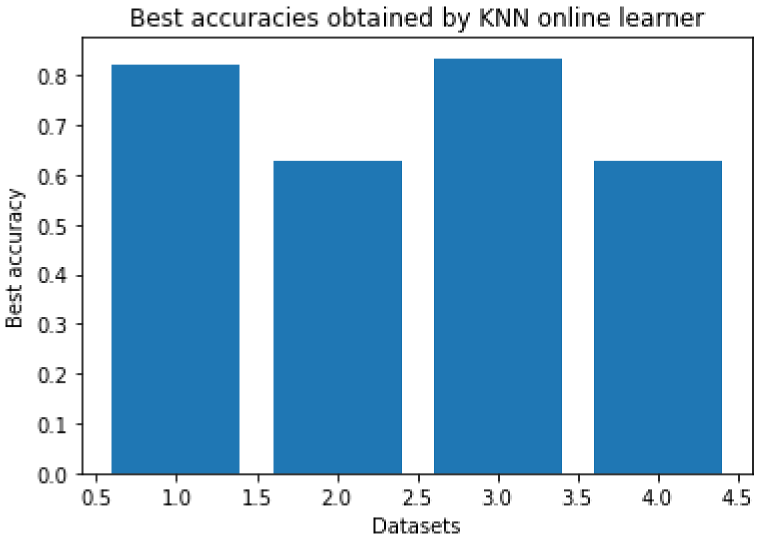

Figure 2. Best accuracies obtained by KNN classifier

c. The best accuracies for all datasets have been obtained when the hyper parameter maximum sample size is 10,000. The best accuracies are: 87.5%, 68.4%, 87.8%, and 67.3% respectively.

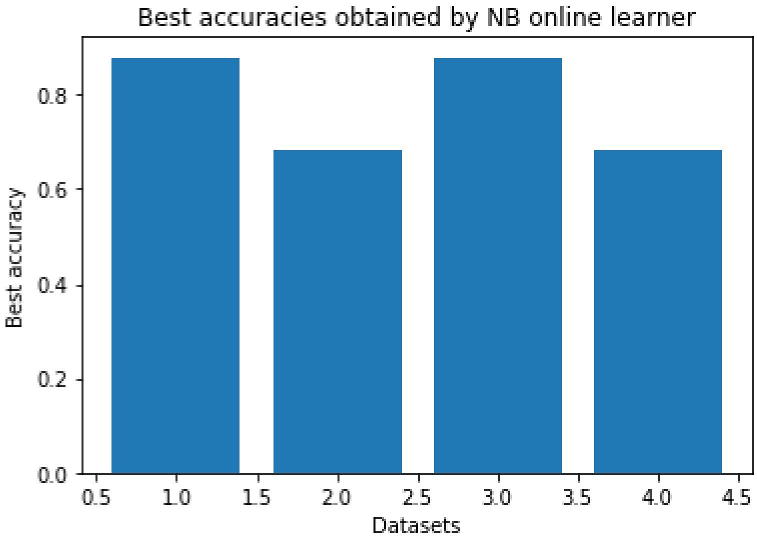

Figure 3. Best accuracies obtained by NB classifier

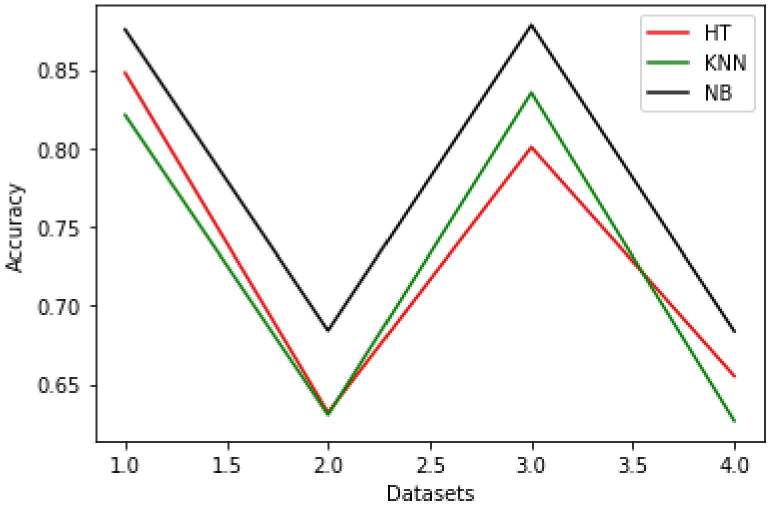

Figure 4. Combined accuracies of the online learners

Question 3

In this question, ensemble classifiers which combine HT, KNN, and NB for the four datasets generated in the first question have been created.

In order to make the model more reliable, the datasets created in the first question which consist of 20,000 samples have been splitted into training and test sets. The ratio of these sets is 80:20 respectively.

a. **The first ensemble classifier uses majority voting rule (MV).**

The accuracies are 88.72%, 70.08%, 88.48%, 67.65%.

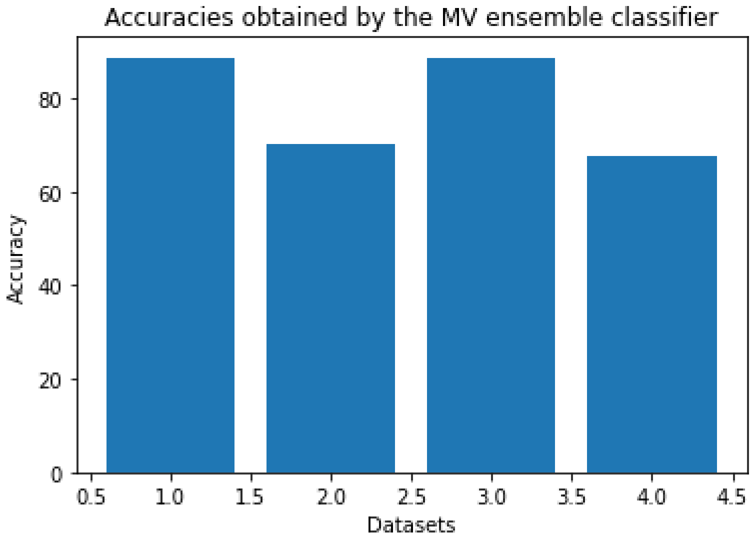

Figure 5. Accuracies obtained by the ensemble classifier with majority voting rule

b. The second ensemble classifier uses the WMV voting rule. In order to determine weights, different types of weights have been used and the weights that are giving the best accuracy have been chosen. When the accuracy of the classifiers is observed, it can be seen that the NB classifier performed better than the other two classifiers, that's why, it's given more weight.

The accuracies are: 89.15%, 70.05%, 88.8%, 68.4%.

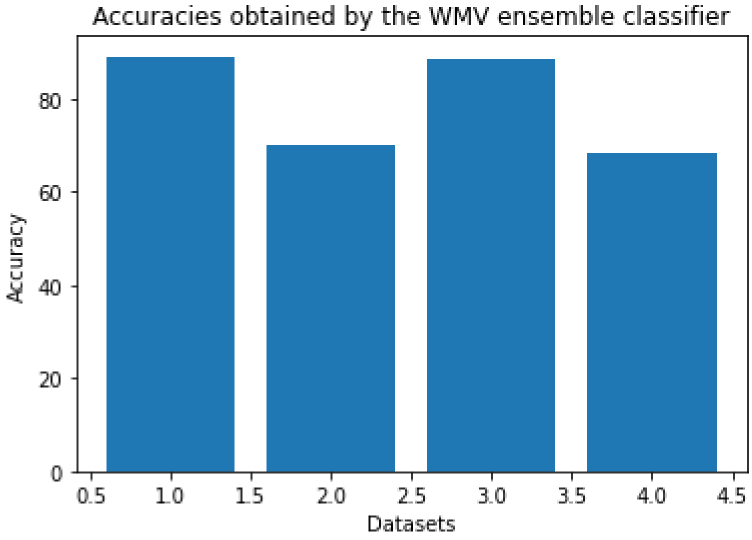

Figure 6. Accuracies obtained by the ensemble classifier with weighted majority voting rule

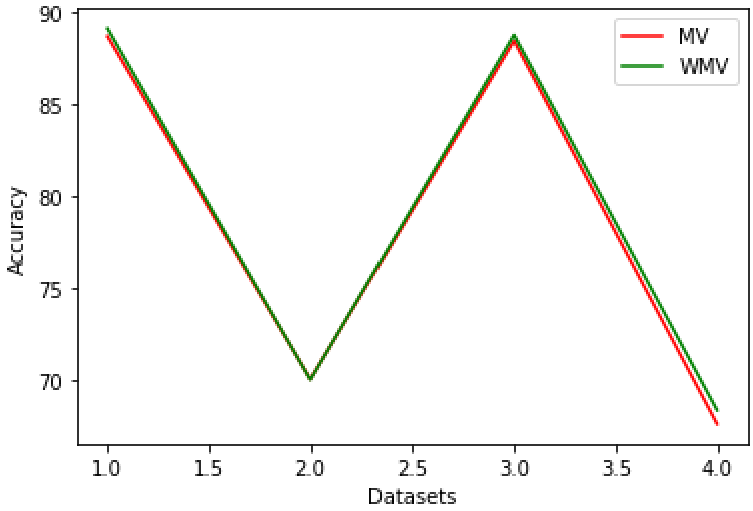

Figure 7. Combined accuracies of ensemble classifiers

Results and Analysis

According to the plots and accuracies that have been obtained, it can be seen that the effect of noise percentage is significant in terms of accuracy. The first and third datasets have noise percentage of 10, and the second and the fourth datasets have noise percentage of 30. Every classifier has achieved more accuracy than the second and the fourth datasets. That’s why, as the noise percentage increases, predicting the label of the sample becomes harder. On the other hand, the effect of drifting features does not have as significant an effect as the noise percentage has.

Elapsed Time and Accuracy Analysis in terms of Batch Size

1. Hoeffding Tree

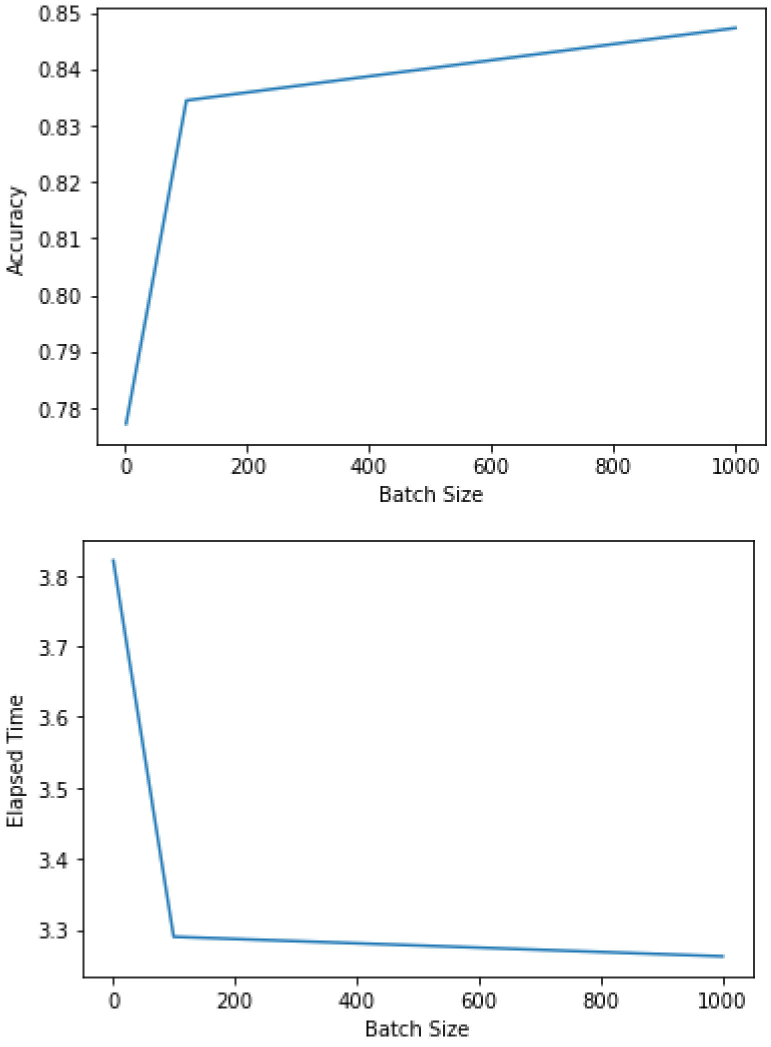

Figure 8. Accuracy and elapsed time plots for the HT classifier

The plots above show the accuracies and elapsed time for the **hoeffding tree** online learner. It can be observed that as the batch size increases the accuracy increases and elapsed time decreases, therefore, it is more logical and efficient to use batch size as 1000 to train the HT online learner.

2. KNN

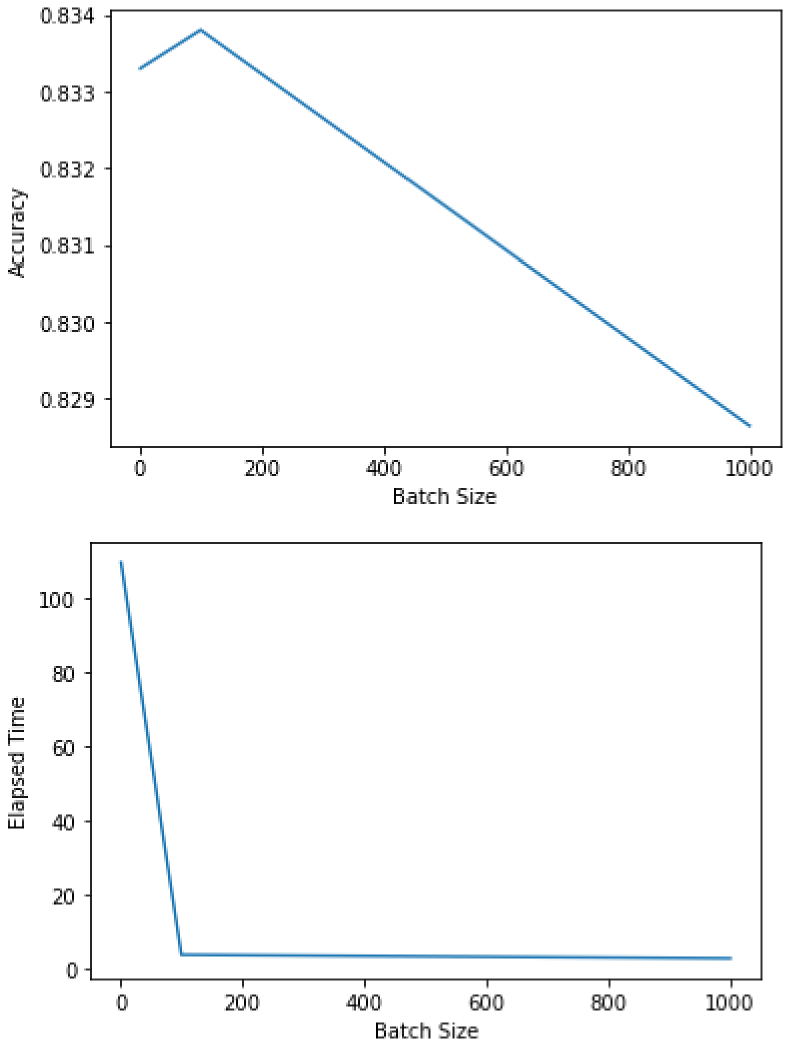

Figure 9. Accuracy and elapsed time plots for the KNN classifier

The plots above show the elapsed time and accuracy in terms of the batch size in **KNN** online learner. It can be observed that when the batch size is 1, the time to train the learner is extremely higher than the other batch sizes. Even though it takes so much time to train a KNN learner when batch size is 1, the accuracy is not higher than the knn learner which is trained when batch size is 1000. Therefore, it is more efficient to use higher batch sizes in KNN.

3. NB

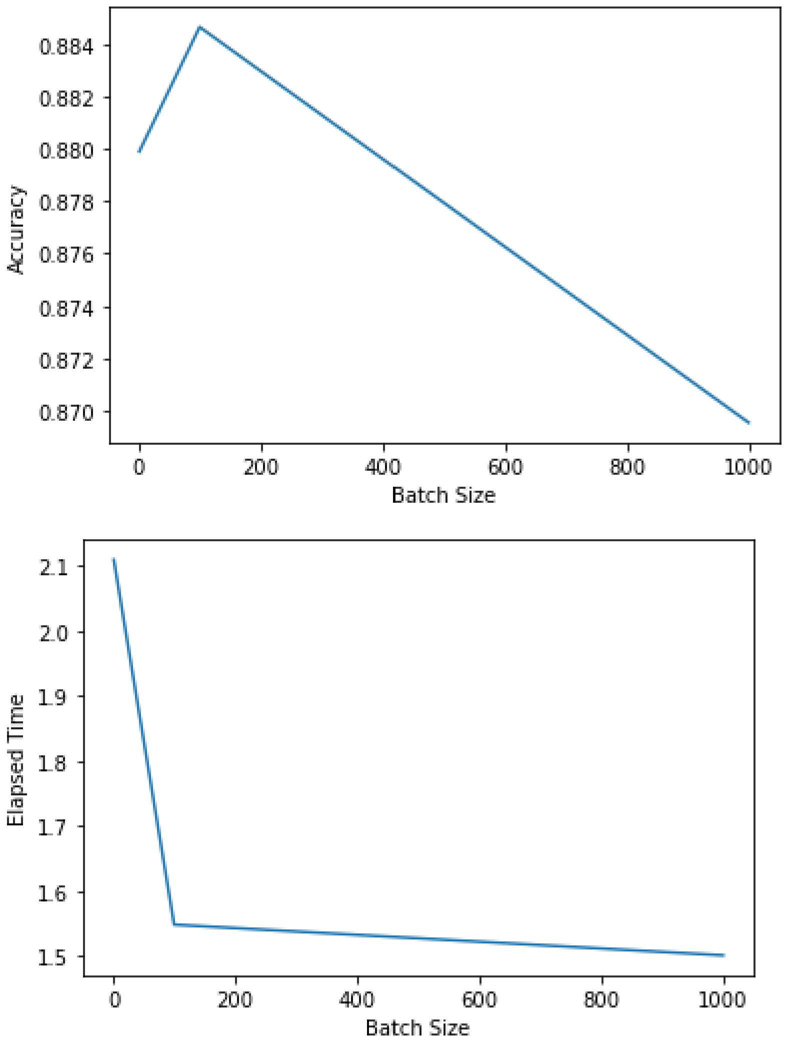

Figure 10. Accuracy and elapsed time plots for the NB classifier

While training NB online classifier, elapsed time is shortest when the batch size is 1000, not surprisingly. Even though the difference between accuracies is negligible, the best accuracy has been obtained when the batch size is 100.

Ensemble Methods Analysis

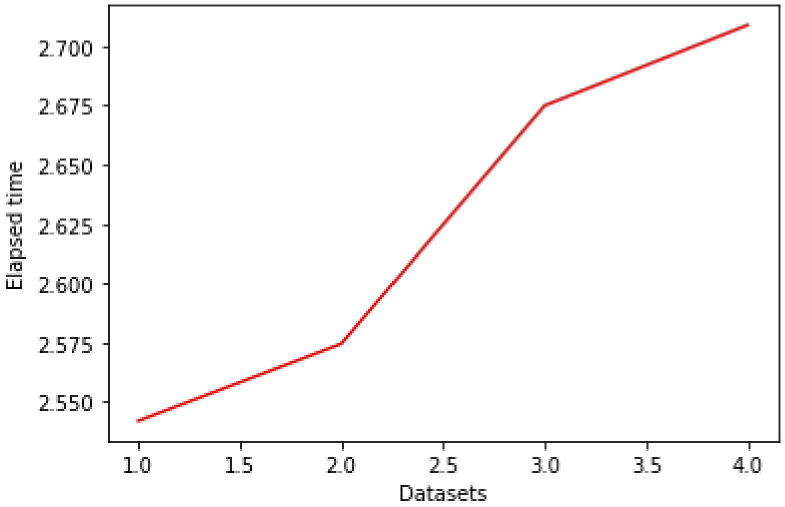

Figure 11. Elapsed time plot for the Ensemble classifier

In the first dataset, the time elapsed to train the ensemble classifier is 2.703 ms which is quite good considering the fact that the dataset contains 20,000 samples.

The accuracies according to datasets are [0.8764, 0.68625, 0.88155, 0.6842], respectively.

The best accuracies obtained by HT: [0.848, 0.63185, 0.80075, 0.65515]
The best accuracies obtained by KNN: [0.8212, 0.6304, 0.8354, 0.6266]
The best accuracies obtained by NB: [0.8756, 0.684, 0.8785, 0.6837]

If the results are observed, it can be seen that the ensemble classifier performed significantly better than KNN. The performances of NB and the ensemble classifier are quite close, however, training NB takes less time than the ensemble classifier.