GE461: Introduction to Data Science
Assignment for Data Stream Mining
Due date: May 15, 2021; 11:59 pm

İlknur Baş 21601847
Ankara, Bilkent University Dormitory, Turkey

## 1. Work to be Done

This section explains the steps that are done in Data Stream Mining assignment and discusses the results.

### 1.1 Dataset Generation

3 SEA Datasets are generated by using SEAGENERATOR from "skmultiflow" library. Creation of these datasets can be seen from the attached python file. Additionally, all datasets have 3 features and a label value for each sample.

### 1.2 Data Stream Classification with Three Separate Online Single Classifiers: HT, KNN, MLP

For the simplicity, all the results are represented as a table

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.9733 |
| noise percentage: 0.1 | 0.90025 |
| noise percentage: 0.7 | 0.3448 |

Table 1: Accuracy result of HT classifier for each dataset

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.97375 |
| noise percentage: 0.1 | 0.8795 |
| noise percentage: 0.7 | 0.60225 |

Table 2: Accuracy result of KNN classifier for each dataset

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.9988 |
| noise percentage: 0.1 | 0.9046 |
| noise percentage: 0.7 | 0.6792 |

Table 3: Accuracy result of MLP classifier for each dataset

### 1.3 Data Stream Classification with Two Online Ensemble Classifiers: MV, WMV

For the simplicity, all the results are represented as a table.

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.9926 |
| noise percentage: 0.1 | 0.90275 |
| noise percentage: 0.7 | 0.67915 |

Table 4: Accuracy result of Majority voting rule for each dataset

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.9874 |
| noise percentage: 0.1 | 0.89525 |
| noise percentage: 0.7 | 0.66895 |

Table 5: Accuracy result of Weighted majority voting rule for each dataset

GE461: Introduction to Data Science
Assignment for Data Stream Mining
Due date: May 15, 2021; 11:59 pm
İlknur Baş 21601847
Ankara, Bilkent University Dormitory, Turkey

## 1.4 Batch Classification with Three Separate Batch Single Classifiers: HT, KNN, MLP

For the simplicity, all the results are represented as a table.

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.9738 |
| noise percentage: 0.1 | 0.9084 |
| noise percentage: 0.7 | 0.3676 |

Table 6: Accuracy result of Batch Classification with HT

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.9984 |
| noise percentage: 0.1 | 0.9038 |
| noise percentage: 0.7 | 0.6868 |

Table 7: Accuracy result of Batch Classification with MLP

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.9798 |
| noise percentage: 0.1 | 0.8832 |
| noise percentage: 0.7 | 0.6214 |

Table 8: Accuracy result of Batch Classification with KNN

## 1.5 Batch Classification with Two Batch Ensemble Classifiers: MV, WMV

For the simplicity, all the results are represented as a table.

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.9862 |
| noise percentage: 0.1 | 0.8902 |
| noise percentage: 0.7 | 0.662 |

Table 9: Accuracy result of Batch Classification with WMV

| Dataset with | Accuracy |
|---|---|
| noise percentage: 0 | 0.995 |
| noise percentage: 0.1 | 0.9034 |
| noise percentage: 0.7 | 0.6848 |

Table 10: Accuracy result of Batch Classification with MV

GE461: Introduction to Data Science
Assignment for Data Stream Mining
Due date: May 15, 2021; 11:59 pm
İlknur Baş 21601847
Ankara, Bilkent University Dormitory, Turkey

## 1.6 Report/Paper: Comparison of Models

In the following lines, comparison for each section is discussed separately. To consider tables between 1 and 3, the accuracy results for dataset with noise percentage 0, are almost identical. However, it can be said that separate online classifier MLP gives the best accuracy result. This is also valid for the dataset with noise percentage 0.1. For each online single classifier, a significant drop occurs in their accuracy results. As it can be seen from Table 1, online classifier HT gives the lowest accuracy result compared to other online classifiers. Table 4 and 5 presents the accuracy results for online ensemble classifiers which are MV and WMV. The difference between the accuracy result of these two tables for each dataset is approximately 0.01. It can be said that both MV and WMV perform better in datasets with noise percentage 0 and 0.1. The dataset with noise percentage 0.7's accuracy result is slightly more than the average yet not good as other datasets. The accuracy results in Section 1.4 are highly similar to the accuracy results in Section 1.2. There is a drastic drop in the accuracy result of the dataset with noise percentage 0.7. In other words, batch classification with HT gives the worst accuracy result with the dataset set of noise percentage 0.7. Same dataset for other classifications is still performed not very accurately but compared to HT classification, it is better. Both batch ensemble classifiers are performed very accurately in dataset with noise percentage 0, and also their accuracy result are very similar as it can be seen from Table 9 and 10. The accuracy results in the tables which are in the Section 1.5 is very similar to one's in the Section 1.3. There is a drastic drop in the dataset with noise percentage 0.7 in terms of its accuracy result. In overall, ensemble methods are performed better compared to individual models especially in the dataset with noise percentage 0.7. It is encountered that, with HT single classifiers the accuracy for this dataset drops to approximately 0.3.