GE461: **Introduction to Data Science**
Assignment for Data Stream Mining
April 25, 2022
Due date: May 11, 2022; Wednesday 11:59 pm
Uploading Your Work: See below

Notes: This assignment is about data stream mining. Classifying data streams is a challenging problem due to time and memory limitations as well as variation in data distribution. Our aim is to effectively classify the data as they continuously keep entering to the system. In your work you will self learn and use scikit-multiflow[1], which is a data mining framework for the Python programming language.

Teaching Assistant: Sepehr Bakhshi, sepehr.bakhshi@bilkent.edu.tr

TA Zoom Office Hours: Between 1 pm and 3 pm on May 7 Saturday. Sepehr will send you the necessary zoom meeting information by email a few minutes before the office hour meeting.

TA Office Hour Administration: Please send your question to Sepehr before his office hours and do it as early as possible for efficiency.

A. What to Submit

Your submission has two components.

1. **Code**. It must contain proper comments. You must also include your name at the top as a signature that confirms that you are the programmer. Please remember that MOSS is in our plans for plagiarism check.
2. **Report**. Your report must be in pdf form and must cover all "Work to be Done" sections of the assignment. For each section of your work briefly explain the purpose and what has been done and achieved in that section. Provide a comparison of results that contains tables and plots as appropriate. Make sure that you follow the principles of scientific writing. Use simple past or simple present tense in your report. If you plan to propose future work then in that case you may use future tense.

   Your report must have proper title like a scientific paper, reflecting its true content. It must have your name and address etc. If you like, for experience and fun, you may use the ACM conference paper format[2]. You must use latex or Microsoft Word or their equivalent.

   Optional: As an optional part, at the beginning of your report, you may have a related works section that covers data stream mining briefly with proper references.

   Optional: Another optional part is comparison of the effectiveness of the methods using statistical tests. The design and administration of these tests should be decided by you by looking at the available papers in literature.

   See Justin Zobel's book *Writing for Computer Science* for further hints on the style of CS related scientific paper writing.

B. Submitting Your Work

You will submit your work by uploading it to Moodle in a zipped file. Its name must be streamMiningYourFirstNameYourLastName. For a student with the name "Ali Can Ok" it is streamMiningAliOk. You may do multiple uploads only the last one will be used for

[1] https://scikit-multiflow.github.io/
[2] https://www.acm.org/binaries/content/assets/publications/taps/acm_layout_submission_template.pdf

C. Work to be Done

Your work has four components.

1. **Dataset Generation**

   Generate Hyperplane dataset with 20,000 instance using Hyperplane Generator. Your dataset should have 10 features and class labels are the default for hyperplane, the other options are modified based on the instructions in part (a-d). It should be something like below data instances. The last column should be the class label.

   0.6987, 0.2568, 0.570, 0.949, 0.1970, … , 0.3285, 0.4474, 0.3355, 0.585, 0.5411, 0
   0.0679, 0.0819, 0.6529, 0.9023, 0.314, … , 0.788, 0.3094, 0.3311, 0.4241, 0.342, 1

   a. **Hyperplane Dataset (noise= 10%, number of drifting features 2)**
   Generate a dataset with 20,000 instances; again 10 features and the other options unchanged; but this time with " Hyperplane Generator" with noise percentage of 10 and number of drifting features equal to two, and write it into a file called "Hyperplane Dataset 10_2".
   b. **Hyperplane Dataset (noise= 30%, number of drifting features 2)**
   All the parameters are like the dataset in (b), except that the noise percentage is 30% and number of drifting features is equal to two.
   c. **Hyperplane Dataset (noise= 10%, number of drifting features 5)**
   All the parameters are like the dataset in (b), except that the noise percentage is 30% and number of drifting features is equal to two.
   d. **Hyperplane Dataset (noise= 30%, number of drifting features 5)**
   All the parameters are like the dataset in (b), except that the noise percentage is 30% and number of drifting features is equal to two.

2. **Data Stream Classification with Three Separate Online Single Classifiers: HT, KNN, MLP**
   Write a script in Python that constructs and trains the following online classifiers using the four Hyperplane Datasets generated in step 1.
   a. HoeffdingTree as **HT** online learner,
   b. K nearest neighbour as **KNN** online learner,
   c. Naïve Bayes as **NB** online learner

3. **Data Stream Classification with Two Online Ensemble Classifiers: MV, WMV**
   Write a script in Python that constructs and trains the following ensemble classifiers that combines HT, **KNN**, and **NB** for the four Hyperplane Datasets generated in step 1.
   a. Majority voting rule **MV,**
   b. Weighted majority voting rule **WMV.**

4. **Report/Paper: Comparison of Models**
   a. In your comparison of models consider the following and additional questions as needed by considering the results you obtained for four datasets.
   b. Compare temporal accuracies of online classifiers, the ones generated in the steps 2, 3 using Interleaved-Test-Then-Train approach using instances of the datasets you generated in the first step. Try to explain the effect of drifting features and noise percentage on the performance of the classifier.
   c. In the comparison of classifiers try different batch sizes (1, 100 ,1000) and discuss if batch sizes are influential in understanding the performance of the methods (in terms of accuracy and runtime). The difference between online and batch mode is that in online mode, data items are processed one at a time (batch size is one). Meaning that, a single data item is first tested and then used for training. However, in batch mode of stream processing, first, a set of data items is collected (based on the batch size) and then this chunk of data is processed altogether. In both of these methods, interleaved

test-then-train approach is used as we are processing a data stream, not a static dataset.

d. Are ensemble methods better than individual models? Explain why.
e. Compare all online and batch models (single and ensemble) in terms of their overall accuracies. How do they compare with each other in terms of overall accuracy and why?
f. How can you improve the prediction accuracy of the online classifiers (single and ensemble)? Try to find a method and incorporate it to your test cases. Try to show it at least with one classifier.

Is there any difference among the models in terms of their efficiency? Does batch size affect the runtime? provide quantitative data about this.

In your comparisons use plots and tables when appropriate. Number all plots and tables and provide proper subtitles for them. Make sure that you refer to each of them in the text of your report. Help your reader by providing a simple and easy to follow presentation.